

nature

THE INTERNATIONAL WEEKLY JOURNAL OF SCIENCE

CHANGING FACE OF PLUTO

New Horizons view
of Sputnik Planum
reveals a self-renewing
nitrogen ice surface

PAGES 40, 79 & 82

SCIENCE POLICY

GOING FOR GOLD

High-spending South Korea
has Nobel prizes in its sights

PAGE 20

PHARMACEUTICALS

INDUSTRIAL REGENERATION

Green chemistry is
good for business

PAGE 27

GEOPHYSICS

EARTH'S CORE MYSTERIES

Extreme experiments put
numerical models to the test

PAGES 45, 95 & 99

NATURE.COM/NATURE

27 June 2016 £10

Vol 534, No 7605



9 770028 083095

22

THIS WEEK

EDITORIALS

FAR SIGHTED Four hundred years of telescopes at a glance **p.6**

WORLD VIEW The developing world can and should do pure science **p.7**



MOONSHINE Chinese home-brew beer kit is 5,000 years old **p.9**

Dark satanic wings

Just as the dark-coloured pepper moth disappears from northern England, researchers are finally getting to the bottom of how it gained its colour.

Not for nothing was the region of the English midlands north of Birmingham called the Black Country during the late nineteenth century. It was the dark polluted heart of the industrial revolution, according to a railway guide from 1851: “The pleasant green of pastures is almost unknown, the streams, in which no fishes swim, are black and unwholesome; the natural dead flat is often broken by high hills of cinders and spoil from the mines; the few trees are stunted and blasted; no birds are to be seen, except a few smoky sparrows; and for miles on miles a black waste spreads around, where furnaces continually smoke, steam engines thud and hiss, and long chains clank, while blind gin-horses walk their doleful round.”

A few kilometres to the north, where trees remained, the wild-life was already adapting to its new human-made environment. You have all seen the results. The first famous dark-coloured peppered moth — a staple of textbooks — was recorded in Manchester in 1848. Half a century later, they were everywhere. The wild-type, light-coloured and mottled, had disappeared almost to extinction.

There is perhaps no better example of natural selection in action than the case of the peppered moth (*Biston betularia*). As the same textbooks say, the colour of the moths evolved to match their new, sooty, backgrounds, and thereby camouflage the insects from hungry birds.

The story is not actually quite so clear-cut. Geneticists have squabbled over the details for decades — the strength of the evidence for the assumed choice of the birds, for example — and some of these technical criticisms have leaked, out of context, into the propaganda of creationists. In response, some textbooks have — heaven forbid — evolved to not include the peppered-moth story at all.

Among the holes in the story was the identity of the gene that was involved in producing the dark-coloured — melanic — moth variant.

Extensive mapping has pinned it down to a 400-kilobase region containing 13 genes, none of which had any obvious role in wing coloration. Undeterred, scientists went on to isolate the gene responsible, and they describe their search on page 102 of this issue. It is called *cortex*, orthologous to a gene of the same name in *Drosophila*. The researchers have even gone further, and shown that the specific cause of the mutation is the insertion of a transposable element (popularly, a ‘jumping gene’) into the first intron of the *cortex* gene.

The insertion leads to increased transcription of the gene during a phase of development when the wing discs are forming. The *cortex* gene, then, is involved in wing development, but there is still no obvious association with coloration. In *Drosophila*, *cortex* is involved in cell-cycle regulation, in particular, marking proteins that are redundant in the cell cycle as being ready for disposal. What is going on?

Work from a different group of Lepidoptera might offer a solution. In a study described on page 106, another group of researchers shows that *cortex* is a key player in the coloration of the wings of butterflies in the genus *Heliconius*, long a favourite for the study of mimicry. They show that *cortex* is a member of a fast-evolving scion of an

otherwise conservative group of cell-cycle regulator genes known as the *fizzy* family, a name redolent of activity, growth and fervour, and possibly involved in the regulation of wing-scale development. This is important, because it is the size, density and surface properties of the wing scales that determine colour in butterflies and moths. Flies, such as *Drosophila*, lack these structures, perhaps explaining why it was initially hard to associate the *cortex* gene with wing development.

“There is enough in the pages that follow to update the textbooks.”

There is a further, satisfying twist to the tale. Although it is possible that melanic mutants existed undetected at a very low level in the peppered-moth population for centuries, the specific mutation behind their coloration is relatively recent, appearing

around 1819 — in plenty of time for it to be noted down in Manchester a couple of decades later.

Much, of course, remains to be discovered, not least of which is the precise mode of action of *cortex*; how the gene relates to wing-scale development; and how the insertion of a transposable element contrives to alter this. But there is enough in the pages that follow to update those textbooks. Still, future generations of readers will find it harder to recognize the high hills of cinders and spoil from the mines that drove the change. The air is cleaner these days, ‘Black Country’ is no longer an apt description, and the dark-coloured peppered moths are vanishing as quickly as they emerged. ■

Toxic control

The United States is overhauling its chemicals law; now it must tackle carbon emissions.

The 1976 US Toxic Substances Control Act (TSCA) must be one of the worst pieces of environmental legislation ever devised. Rather than empowering the Environmental Protection Agency (EPA) to ensure that new chemicals are safe, the law declared all chemicals harmless, unless proven otherwise. The situation is so preposterous, in fact, that even the normally dysfunctional US Congress managed to unite last week to advance reform (see page 18).

The bipartisan TSCA reform bill passed the House of Representatives, by a vote of 403–12, on 24 May. Although senator Rand Paul (Republican, Kentucky) has temporarily blocked a vote in the Senate, the legislation is expected to pass in the coming weeks, clearing the way for a signature by President Barack Obama. Once that happens, EPA scientists will at last have the authority to do their jobs.

Rather than watching passively as some 700 new chemicals enter all corners of the US marketplace each year, the EPA would be able to require companies to provide more data and conduct extra research to demonstrate the safety of the products. The legislation would also bolster review of existing substances. The TSCA inventory currently lists some 85,000 chemicals, but no one knows how many are still in use today. The EPA would create a new inventory and then sift through it to see which ones merit further investigation.

What is most remarkable about this reform legislation — aside from the fact that it took so long — is the list of supporters: Democrats and Republicans, both houses of Congress and the legislative branch, as well as many environmentalists and the chemical industry. The reason is simple: the companies that manufacture and use chemicals, once adamantly opposed to such reform bills, have realized that a viable federal regulatory system is in their financial interest. The complete lack of public confidence in the EPA's authority under the TSCA has pushed environmental officials at the state level to launch their own investigations and regulations. The upshot is that without a stronger federal system, the industry faces an increasingly complex — and uncertain — patchwork of regulations.

This is all good news for the public, which is bombarded daily by news reports, environmental campaigns and scientific studies that analyse the danger of one chemical or another in products that they purchase every day. It is also good for science. The new law will drive research into chemicals of concern, and companies will find it harder to claim that the information that they submit is a trade secret. As a result, more data will enter the public and academic spheres, and that is always a good thing.

Environmentalists pushed to ensure that the EPA's new decisions about health risks will be based on health data alone, without regard to economic implications. Under the new legislation, the EPA would be able to consider economic impacts in any subsequent cost-benefit

analysis only if it moves forward with regulations. And industry pushed for mandatory deadlines to ensure that decisions are made in a timely manner. All in all, it's a reasonable compromise that moves the regulatory needle in the right direction.

It is also a blueprint for what ultimately needs to happen to break the legislative stalemate on what is perhaps the greatest environmental challenge: the effect of greenhouse gases on climate.

"It's a reasonable compromise that moves the regulatory needle in the right direction."

Despite overwhelming evidence showing the need for action, the energy industry has obstructed and stalled for too long, and the only real result is prolonged regulatory uncertainty. If major businesses, including energy producers and consumers, were to get together en masse and push for regulation, Republican lawmakers would be forced to pull their heads out of the sand

and think about reasonable solutions that are in line with their own political values.

Low-carbon energy such as nuclear power and that obtained from renewables would benefit the most, but natural gas would also get a short-term boost as utilities back further away from coal, which is already on the decline. Even coal would see its chances of survival increase in the long run, because properly agreed federal regulations would bolster the economics and interest in technologies that can be used to capture and sequester, or even use, carbon dioxide. At a minimum, with a legitimate set of rules in place, companies could move forward and plan their long-term investments accordingly.

Everyone could see that the original TSCA bill created a problem. It has taken decades, but reform was inevitable. The need for legal controls on the generation and control of greenhouse gases is just as clear — indeed, that is why the energy industry has fought so hard to undermine the evidence. This time, we do not have decades to waste. ■

Seeing farther

Our fascination with telescopes and the worlds they reveal spreads beyond science into culture.

Galileo Galilei did not invent the telescope, but he is generally credited with being the first to point one at the sky and record what he saw. Which begs a question: just what did the others before him do with theirs?

Ever since the great man saw and drew the moons of Jupiter in 1610, astronomers — both amateur and professional — have been captivated by the night sky. For more than 400 years, through revolution, war and endless change on Earth, telescopes have brought the rest of the Universe to us in ever-greater detail. We perch them on the tops of the highest mountains, strap them to aircraft, dangle them from balloons and launch them into orbit, all to get a better view of the world outside our own. We even cut holes in the roofs of our houses for them. The word 'telescope' derives from the Latin for far-seeing, and never can a scientific instrument have been so well labelled.

On page 34, Bernie Fanaroff, who as the former head of the Square Kilometre Array South Africa project knows a thing or two about telescopes, reviews a new account of their development and history. *Eyes on the Sky* by Francis Graham-Smith covers the entire spectrum, from existing instruments to planned ones that gather everything from long-wavelength radio waves to high-frequency X-rays. Readers with a taste for the bizarre could also check out *Unusual Telescopes* by Peter Manly, published in paperback in 1995. Among the weird and wonderful designs are telescopes with mirrors made from polished rock, inflatable telescopes and ornamental telescopes

that double as sundials.

The names of some of the newest additions to the telescope roster — some barely off the drawing board — indicate where the field is heading. The Very Large Telescope will soon be joined by the European Extremely Large Telescope, but not by its cancelled rival, the Overwhelmingly Large Telescope.

But small instruments can be powerful, too, if there are enough of them. Maybe the future of astronomy lies not in ever-bigger adverbs but in tiny chips: a News story on page 15 offers a glimpse of that perhaps-not-too-distant technology. Next month, a package that holds dozens of Sprite mini-satellites is scheduled to be sent to the International Space Station, from where they will be released. It is a test run to gauge the potential of such 'chipsats' to swarm and collectively gather data on missions.

Next month will also see a telescope-related launch of a different kind — a new festival at the historic UK Jodrell Bank observatory near Manchester, headlined by the French musician Jean Michel Jarre. It was scientists at Jodrell who famously, with the help of a fax machine borrowed from the *Daily Express*, scooped the Soviets and intercepted the first pictures of the lunar surface from the Luna 9 mission. The glory days of that observatory may be behind it, but its status as an iconic landmark demonstrates another feature of telescopes: they provide a tangible link not just from astronomers to the Universe but from science to the wider public, especially when it involves an enormous radio dish. Indeed, the United Kingdom is seeking to have the site's cultural significance marked officially: Jodrell Bank is being considered for listing as a UNESCO World Heritage Site.

Telescopes and their discoveries have always spread beyond science. Shortly after Galileo drew Jupiter and its four moons, William Shakespeare is thought to have completed *Cymbeline*, one of his final plays. At its climax, the god Jupiter descends to the stage, preceded by four angels. Science and culture have never looked back, or so far. ■



The developing world needs basic research too

The establishment of an agency in Indonesia that will support 'frontier research' is a welcome development, argues Dyna Rochmyaningsih.

What use is basic science to the developing world? Why would a nation that cannot feed all its people try to send a spacecraft to Mars? Instead, scientific research in poorer nations is expected to focus on applied problems. Surrounded by poor prospects and infrastructure, institutions in these countries support fast-producing research that can provide direct results to the economy.

Much foreign investment goes the same way. For developing countries to work on pure science is often viewed from outside as indulgent and wasteful. Witness the argument that took place in the United Kingdom last year over India — a recipient of British aid — developing its own space programme.

Applied research certainly has its place, in developing as well as developed countries. Science as a tool to make money and secure a food supply is key to survival. It can target local issues: a notable success is the process developed by Brazil to turn (locally abundant) sugar cane into ethanol as a biofuel resource. In southeast Asia, the science of cassava pests and diseases is a priority, because millions of people here rely on cassava as a staple food and a source of income.

But we should not forget that there is more to life than accumulating resources. Many other factors threaten human existence — from mutating viruses to moving tectonic plates. In the Southern Hemisphere, where most developing countries are located, natural disasters and emerging diseases haunt the lives of billions of people. The Ebola epidemic and the Zika virus, the Ecuador earthquake and the Aceh tsunami are just a few recent examples. And to understand them, we do not just need science that has an economic value. We need science that questions why the world is the way it is.

Of course, some in the developing world already study pure science problems. In Indonesia, some researchers are analysing the genetics of Indonesian people and their susceptibility to certain diseases — work that also offers insights into human origins. Others are studying the ecology and evolution of non-human primates. But these efforts are dwarfed by the many government-funded projects on applied topics such as agriculture, pharmacy and animal husbandry.

Besides the fact that it has less economic value, basic science is not encouraged in developing countries because it is expensive. Almost all such countries allocate less than 1% of gross domestic product to scientific research. In 2016, the grant from Indonesia's Ministry of Research and Technology for a research project rarely exceeded US\$100,000 — not enough to buy cutting-edge laboratory equipment. We see a similar picture in other developing countries, including many in Africa.

Things are starting to change. Earlier this year, President Joko Widodo of Indonesia signed into existence the Indonesian

Science Foundation (ISF), an independent funding body for science. The establishment of the ISF is a monumental event. For the first time, Indonesian scientists will have a funding source apart from the national budget (of which the proportion going to science is a very low 0.08%). And, also for the first time, they will get multi-year research grants. The amount will be increased, up to \$300,000 per successful research proposal. As a start, the Ministry of Finance has committed to provide \$9 million in 2016 for research on life sciences, health and nutrition.

And the most interesting part is that the new funding agency will not support applied science. Instead it will pay for 'frontier research' on the Universe, Earth, climate, the life sciences, health, nutrition, materials and computational science.

The new programme might encourage the best Indonesian scientists scattered across the developed world to come back. It should encourage those in Indonesia to do better science. It will certainly grow scientific excellence in the country. Unlike applied science, the goal is not to use research as a tool, but for it to become a valuable and self-sustaining pursuit in its own right. The ISF is intended to create a system in which scientists can work independently, without the need for international support, to assess the scientific questions of their own land and to contribute to the universal quest for knowledge. It offers an opportunity for our scientists to stand on their own feet.

The importance of basic science in poorer countries is recognized beyond Indonesia.

Earlier this year, at a meeting to promote scientific talent in Africa, Mary Teuw Niane, minister of higher education and research in Senegal, spoke of the need for basic science in his and other developing nations.

The African Academy of Science is working with funders including the Wellcome Trust and the Bill & Melinda Gates Foundation to boost basic research in health care. Last month, some £21 million (\$31 million) was awarded to scientists from Côte d'Ivoire, Kenya, Senegal and Uganda who are conducting research on emerging infectious diseases, neonatal and population health, and the elimination of malaria.

It is too early to make predictions, but perhaps we can be optimistic that a new focus on basic research will produce a lasting change in science in the global South. Basic science may not give us an instant result but it will give us a deeper understanding about the world that changes all the time. And it will generate knowledge, which as policymakers from across the world insist, is at the heart of the modern economy. ■

Dyna Rochmyaningsih is a freelance science journalist in Jakarta.
e-mail: drochmya87@gmail.com

WE NEED
SCIENCE
THAT
QUESTIONS
WHY THE
WORLD
IS THE WAY IT IS.

RESEARCH HIGHLIGHTS

Selections from the
scientific literature

ASTRONOMY

Galaxy from the cosmic dark ages

Astronomers have found the faintest example yet of a galaxy from the early Universe.

Kuang-Han Huang of the University of California, Davis, and his colleagues spotted the 13-billion-year-old galaxy using the Keck Observatory in Hawaii and the Hubble Space Telescope. A cluster of galaxies in between acted like a lens, gravitationally bending light from the faint galaxy to make it visible to the telescopes.

The detected galaxy is from the end of the 'cosmic dark ages' — when ultraviolet radiation from the earliest stars ionized the Universe's hydrogen to generate the levels seen today. The authors say that studying more galaxies like this one could reveal whether stars did this alone or had help from other sources, such as black holes.

Astrophys. J. Lett. 823, L14 (2016)

METABOLISM

Growth factor treats diabetes

Injecting a protein into rodent brains triggers long-term remission of type 2 diabetes.

Certain types of protein called fibroblast growth factors (FGFs) decrease blood glucose levels when they are injected into the bloodstream of animals. To see whether they target the brain, Michael Schwartz of the University of Washington in Seattle and his colleagues injected the brains of rats and mice that had type 2 diabetes with one-tenth of the amount of FGF1 used for bloodstream injections. They found that blood glucose decreased to normal levels 7 days after injection, and

stayed that way for up to about 4 months. FGF1 did not change body weight, food intake or blood insulin levels, but glucose was cleared from the circulation into the liver and skeletal muscles twice as fast in treated mice as in untreated ones.

Brain injection of FGF1 may combat diabetes by regulating neural circuits that control how the liver takes up glucose after meals, pointing the way towards possible drug targets, the authors speculate.

Nature Med. <http://dx.doi.org/10.1038/nm.4101> (2016)

PLANETARY SCIENCE

End of a Martian ice age

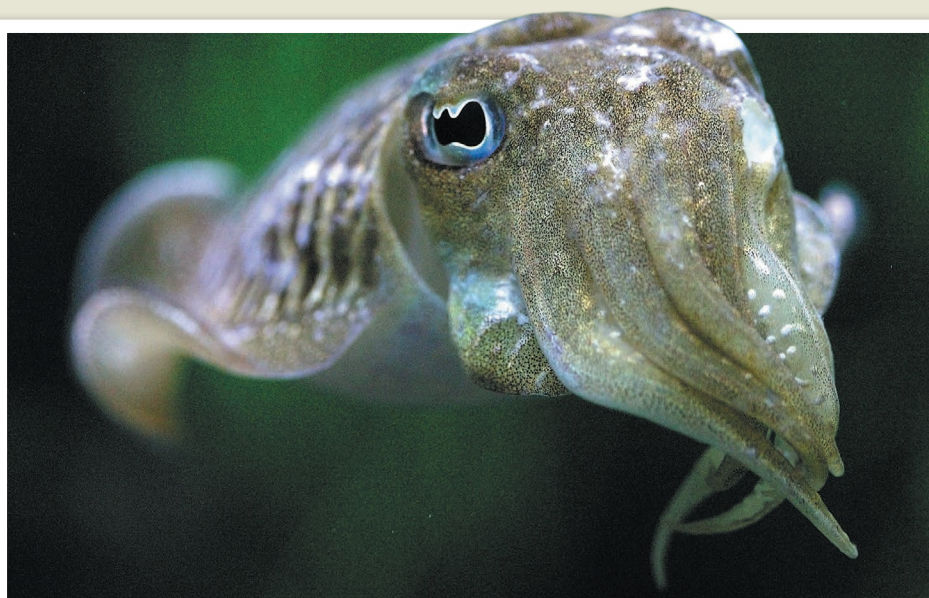
Frosty layers at Mars's north pole show that the planet is emerging from an ice age.

Mars experiences big climate shifts because of the way it tilts on its axis and orbits the Sun. Isaac Smith of the Southwest Research Institute in Boulder, Colorado, and his colleagues used a radar instrument aboard the Mars Reconnaissance Orbiter spacecraft to hunt for signs of

these changes at the north pole. The geometry of the ice layers — sometimes flat, sometimes cutting across one another — allowed the scientists to work out the history of the ice. Some 87,000 cubic kilometres have built up since the end of the last ice age about 370,000 years ago.

The researchers conclude that the ice age is ending because most of this ice accumulated at the north pole, which, unlike on Earth, is warmer than the rest of the planet during an ice age.

Science 352, 1075–1078 (2016)



GERARD LACZ/REX/SHUTTERSTOCK

POPULATION ECOLOGY

A boom in octopuses and cuttlefish

Cephalopods, such as squid, cuttlefish and octopuses, may be benefiting from changes to their environment.

Zoë Doubleday and Bronwyn Gillanders at the University of Adelaide in Australia and their colleagues compiled data from fisheries and scientific marine surveys on global cephalopod catch rates since 1953. They found that cephalopod populations (pictured is a *Sepia* cuttlefish species) have increased significantly over the past 60 years across some 35 species

with different lifestyles, such as ones that live on the sea floor and in the open ocean.

Cephalopods have short lifespans, rapid growth rates and are highly adaptable, which in changing conditions (such as ocean warming) could give them an advantage over slower-growing organisms, the authors say. The cephalopod boom, however, could have damaging effects on their prey populations, such as certain fish and marine invertebrates.

Curr. Biol. 26, R406–R407 (2016)

CELL BIOLOGY

How prions kill brain cells

Brain-wasting proteins called prions kill neurons by shortening the dendritic spines that the cells use to transmit signals to each other.

Prions are infectious and cause neurodegenerative diseases such as scrapie in animals and Creutzfeldt–Jakob disease in humans. To learn how they kill brain cells, David Harris at Boston University in Massachusetts and his co-workers exposed cultured mouse neurons to the prion that causes scrapie in mice. They found that the neurons' dendritic spines retracted within 24 hours, before the cells died. This occurred only in neurons that made the normal, non-infectious form of the prion protein, which suggests that the disease-associated prion might bind to the normal one to trigger dendritic loss.

This method could be used to test potential drugs against prion diseases, the authors say. *PLoS Pathog.* 12, e1005623 (2016)

ARCHAEOLOGY

Ancient beer recipe from China

A 5,000-year-old brewery in China used what was then an unusual ingredient — barley.

A team led by Jiajing Wang at Stanford University in California analysed starch grains from pottery resembling brewing vessels (reconstructions pictured), which were discovered at the Mijiaya site in northern China about a decade ago. The vessels contained a mixture of millet, tubers, a tropical grass known as Job's tears, and barley. Some grains were swollen and deformed as though they

had been mashed, a process that uses hot water to extract sugars. Chemical analysis of residues on the pottery revealed calcium oxalate, a common by-product of beer making.

Barley was domesticated in Western Eurasia around 10,000 years ago, but it did not become a major crop in China until around 2,200 years ago. The Mijiaya brewers may have seen barley as an exotic treat, the authors suggest.

Proc. Natl Acad. Sci. USA
<http://doi.org/bhwm> (2016)

ANIMAL BEHAVIOUR

Onlookers boost mouse chatter

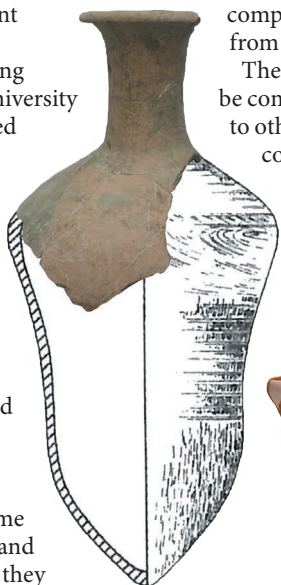
Male mice communicate more in front of an audience than when they are alone.

Mice live in large social groups and communicate using ultrasonic frequencies. To learn how this social environment influences their vocalizations, Roian Egnor of the Howard Hughes Medical Institute's Janelia Research Campus in Ashburn, Virginia, and her colleagues exposed male mice in the lab to a female odour. They then compared vocal responses from animals that were alone to those that were in the presence of another male. Males that had an audience produced vocalizations that

were longer and more complex than those from solo males.

The male mice could be communicating to other males to compete for mates, the authors suggest.

J. Exp. Biol. 219, 1437–1448 (2016)



20 cm

ZOOLOGY

Squid may reach epic sizes

Giant squid could measure up to 20 metres in length.

Fewer than 500 of the mysterious invertebrates (genus *Architeuthis*) have ever been measured. To calculate their maximum possible size, Charles Paxton of the University of St Andrews, UK, compiled recorded measurements of giant squid, and established relationships between measurements such as the length of the whole body, the mantle and the beak. These relationships, and the variation between observed squid lengths and beak sizes, suggest that the animals could plausibly be 20 metres long.

Paxton says that some giant squid may grow too large to be eaten by some of their predators, such as female sperm whales.

J. Zool. <http://doi.org/bhwq> (2016)



commercially important plant, the authors say.

Nature Plants <http://doi.org/bhwn> (2016)

MATERIALS

Light heals defects in solar-cell film

Intense light shining on a material used in experimental solar cells can improve its performance.

Perovskite films promise to increase the efficiency of solar cells, but imperfections in the material, called traps, limit further gains. A team led by Samuel Stranks of the Massachusetts Institute of Technology in Cambridge found that intense light reduces the density of the traps by tenfold, boosting performance. Chemical imaging revealed that iodine ions migrated away from the illuminated areas, and the authors suggest that this effectively swept the traps away.

The effect fades over time, but the authors hope to devise a method with longer-lasting effects for commercial applications.

Nature Commun. 7, 11683 (2016)

GENOMICS

Genetic clues to more rubber

The genome of the rubber tree has revealed a group of genes that may drive the plant's unique ability to produce vast amounts of rubber.

Scientists had previously released a draft sequence, but Chaorong Tang at the Chinese Academy of Tropical Agricultural Sciences in Danzhou and his colleagues now report a more complete genome sequence for the plant (*Hevea brasiliensis*; pictured). Four members of the *REF/SRPP* gene family, which are thought to be involved in rubber synthesis, were among the most highly expressed genes in latex, the white fluid from which natural rubber is obtained. The researchers also identified more than 500 genes that respond to ethylene, a plant hormone known to stimulate rubber production.

These findings could help to guide efforts to breed higher-yield versions of the

➔ **NATURE.COM**

For the latest research published by Nature visit:

www.nature.com/latestresearch

SEVEN DAYS

The news in brief

POLICY

ITER improvements

The nuclear-fusion project ITER has improved its performance and management, and the United States should continue to support it at least until 2018, the US Department of Energy said in a report released on 26 May. ITER is a collaboration between the European Union, China, India, Japan, South Korea, Russia and the United States. Its goal is to show that fusing hydrogen nuclei to make helium is a feasible way to produce electricity. The multibillion-euro experiment is under construction in southern France, but the work is more than a decade behind schedule, and its costs have spiralled. See page 16 for more.

Science for all

Ministers from the European Union's 28 member states have agreed that open access to scientific publications should become the common standard across the bloc by 2020. The EU Competitiveness Council, which met in Brussels on 26–27 May, announced the target following a public debate of broader plans to develop

NUMBER CRUNCH

263,211

The number of extra deaths from cancer during the financial crisis of 2008–10 in countries that are members of the Organisation for Economic Co-operation and Development. Countries with universal health-care systems seemed to be protected from this impact, according to a study in *The Lancet*.

Source: M. Maruthappu et al. *Lancet* <http://doi.org/bhzz> (2016)



SIMON DAWSON/BLOOMBERG/GETTY

Better barley boosts Ethiopian brewing

Two high-yielding varieties of malt barley might help Ethiopian smallholders. The strains can produce yields of up to 6 tonnes per hectare — triple that of the average traditional crop (pictured). They were released on 26 May by the Holetta Agricultural Research Center near Addis Ababa, after decades of

collaboration with the International Center for Agricultural Research in the Dry Areas, headquartered in Lebanon. Demand for the crop — for food and for Ethiopia's burgeoning beer industry — is outstripping supply, with shortages in 2015 forcing some breweries to cut production.

'open science.' The aim is to make research and data more freely available to scientists and to the wider society. The meeting's conclusions held few specific details as to how the target might be reached, but prioritize open access on the EU political agenda.

Chemical reform

The US House of Representatives approved a historic bill on 24 May, strengthening oversight of both new and old chemicals. The bipartisan legislation would overhaul the 1976 Toxic Substances Control Act — widely considered ineffective — and expand the US Environmental

Protection Agency's authority to ensure that chemicals are safe. The bill, which comes with endorsements from the White House and industry, and cautious support from many environmental groups, is expected to pass the Senate soon. See pages 5 and 18 for more.

FUNDING

French cuts

The French government has backtracked over part of a plan to cut €256 million (US\$285 million) from this year's research and higher-education budget after 8 eminent French scientists called the plan "scientific and

industrial suicide". In response, President François Hollande promised on 30 May to reduce the proposed cuts — 1.1% of the total research and higher education budget — by €134 million. High-profile research agencies — the Alternative and Atomic Energy Commission (CEA), the National Centre for Scientific Research (CNRS), the National Institute for Agricultural Research (INRA) and the computer-science institute Inria — will be spared from cuts, he said.

Telescope record

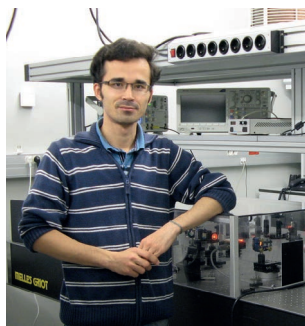
The European Southern Observatory signed a €400-million (US\$448-million)

contract — the largest ever for a ground-based telescope — on 25 May for construction of the dome and structure of the European Extremely Large Telescope (E-ELT). Building is expected to begin in 2017 on the 3,000-metre-high Cerro Armazones peak in Chile, and should be completed in 2024. With a primary mirror 39 metres in diameter and a footprint the size of a football pitch, the E-ELT will be the largest optical telescope on Earth. Construction for the slightly smaller Thirty Meter Telescope in Hawaii has been stalled by protests by Native Hawaiians.

PEOPLE

Iranian physicist

Omid Kokabee, a laser physicist who has been in an Iranian jail for more than five years for “communicating with a hostile government”, has been bailed on temporary medical leave following surgery for kidney cancer, sources tell *Nature*. The 33-year-old scientist, who had studied at the University of Texas at Austin, left a hospital in Tehran on 25 May after his friends posted bail of 5 billion Iranian rials (US\$165,000). They hope to extend his leave using an article of Iran’s penal code that permits the postponement of a sentence that may harm a prisoner’s health. Kokabee (**pictured**)



was arrested in Iran in 2011, while visiting family, and was sentenced to 10 years in prison for alleged espionage — which he denies. Numerous appeals have been made for his release by scientific and human-rights organizations. See go.nature.com/xp4vza for more.

Fossil schemes

Donald Trump, the presumptive Republican nominee for the forthcoming US presidential elections, promised on 26 May to roll back environmental regulations, promote domestic fossil-fuel production and pull the United States out of the 2015 Paris climate agreement if elected. Speaking in North Dakota, where the oil boom has collapsed owing to low oil prices, Trump accused the administration of President Barack Obama of using “totalitarian tactics” and implementing “draconian climate rules” to halt the use of fossil fuels. A global-warming sceptic, Trump said that his

administration would deal with “real environmental challenges, not phony ones”.

Next Science editor

Jeremy Berg will become the editor-in-chief of the Science family of journals, the American Association for the Advancement of Science (AAAS) announced on 25 May. He will succeed Marcia McNutt when she leaves on 1 July to start her role as president of the US National Academy of Sciences. Berg is currently associate senior vice-chancellor for science strategy and planning in the health sciences at the University of Pittsburgh, Pennsylvania. He is a former director of the US National Institute of General Medical Sciences, and will be the 20th holder of the AAAS post.

RESEARCH

Phone doubts

The preliminary findings of a huge animal study are fuelling ambiguity over possible health risks from mobile-phone use. In partial findings uploaded to the bioRxiv preprint website on 27 May, researchers with the US\$25-million US National Toxicology Program (NTP) report that up to 3% of male rats that were exposed to levels of radiation higher than most phone users would experience developed malignant brain and heart tumours (M. Wyde

COMING UP

5–9 JUNE

Astrophysicists and science historians ponder the Science of Time — past, present and future — at a meeting in Cambridge, Massachusetts.
go.nature.com/mhsft6

6–10 JUNE

The biennial Conference on Mathematical Geophysics takes place in Paris. The meeting focuses on experimental investigation as well as theoretical and modelling work.
go.nature.com/lwekvs

et al. Preprint at bioRxiv <http://doi.org/bjfm;2016>). The NTP plans to release data from a similar mouse study in 2017. Whether the final results of the studies may be relevant to humans is unclear. Peer-reviewed studies have previously found no cancer risk associated with mobile-phone use in humans.

EVENTS

Space inflation

Astronauts on the International Space Station successfully tested a flexible orbital habitat on 28 May. The Bigelow Expandable Activity Module, which inflates from 1.7 metres to 4 metres long, is meant to provide an extra 16 cubic metres for living and working in deep space. The module had some initial problems expanding, owing to unexpected friction between layers of its fabric, but was eventually brought to a pressure equal with that of the rest of the space station. The module will remain in orbit for two years, and serve as a test for possible bigger versions in the future.

NATURE.COM

For daily news updates see:
www.nature.com/news

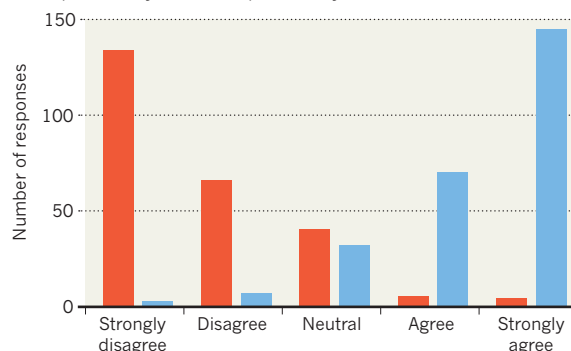
TREND WATCH

A survey of almost 250 biomedical scientists suggests that 80% feel that preprint servers — on which manuscripts are posted before peer review and formal publication — should not be hosted by for-profit organizations. The poll was conducted ahead of a 24 May workshop convened by ASAPbio in Bethesda, Maryland (see go.nature.com/wtkqls). The group is coordinating efforts to implement preprint servers for biology, including making plans for their governance.

PREPRINTS NOT PROFITS

A survey by ASAPbio suggests that the biology community strongly favours the non-profit model for hosting preprints.

A preprint server should be hosted by a:
■ For-profit entity ■ Non-profit entity



NEWS IN FOCUS

PUBLISHING Innovative journal *eLife* gets fresh tranche of cash **p.13**

SPACE 'Chipsat' launch tests alternative way to explore Solar System **p.14**

EXPLAINER Why expansion of US chemical regulation matters **p.18**

EPIDEMIOLOGY High-speed video shows how far sneezes really spread **p.24**

DAVID DOUBILET/NGS



The health of coral reefs is normally assessed by scuba surveys and other close-up views.

OCEANS

Reefs mapped from above

Satellites and research aeroplanes could offer a better, broader view of coral health.

BY ALEXANDRA WITZE

Eric Hochberg has studied coral reefs for two decades, but the marine ecologist is about to see them in a fresh light. Beginning on 6 June, Hochberg and his colleagues will use a specially outfitted NASA aeroplane to map the spectra of sunlight reflecting off reefs spread across the Pacific Ocean far below. The scientists aim to tease out the spectral signatures of coral, algae and sand — and to check the health of the reefs.

The three-year, US\$15-million Coral Reef Airborne Laboratory (CORAL) project will be the biggest and most detailed study yet of entire reefs, rather than just the small patches that scuba divers can reach. CORAL is part of a growing push to map reefs faster, and in more

detail, than ever before. Marine scientists are putting new instruments onto planes, satellites and even drones to gain a broader perspective on how well corals are doing — or not.

After its surveys in Hawaii, Australia's Great Barrier Reef, the Mariana Islands and Palau (see 'Under the sea'), CORAL will have mapped about 3–4% of the world's reef area, hundreds of times more than previous scuba surveys.

Warming ocean waters have led to massive coral-bleaching events such as the one now devastating the Great Barrier Reef. The CORAL scientists hope to learn how individual reefs respond to such threats. "We want to start looking at things at the ecosystem scale, which is really hard to do in the water," says Hochberg, at the Bermuda Institute of Ocean Sciences in St George's.

Remote sensing of coral reefs is hard because the oceans reflect so much less light than the land, says Heidi Dierssen, a marine ecologist at the University of Connecticut Avery Point in Groton, who is part of the CORAL team. And scientists have to do elaborate calculations to correct for the distortion of light on its journey through the atmosphere and through water — a bright, deep ocean bottom and a dark, shallow bottom can both look the same to a remote-sensing camera.

Teasing out such distinctions requires scanning an area using as many wavelength bands as possible. "When you have the full spectrum, you can say so much more about what is there," Dierssen says.

One of the latest views from above comes from the Sentinel-2 satellite, launched by the ▶

► European Space Agency in June 2015. Although the satellite was not designed to study reefs, it has relatively sharp vision and can operate over more and narrower spectral bands than the US Geological Survey's Landsat-8 satellite, another workhorse of Earth observing. And unlike data from keen-eyed commercial satellites, Sentinel's observations are free to use.

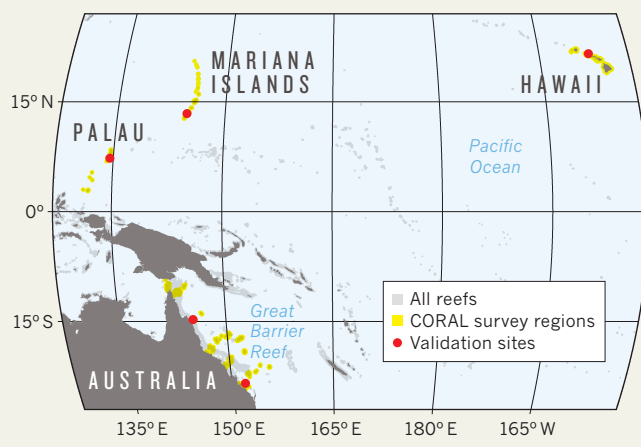
Sentinel-2 will also eventually revisit the same spot every 5 days, compared with Landsat-8's 16-day return period. That makes it a better choice for studying short-term marine phenomena such as coral bleaching and algal blooms, says John Hedley, a remote-sensing expert at Environmental Computer Science in Tiverton, UK, who is on the science team for the Sentinel-2 coral study, Sen2Coral.

Team members are set to report early results on mapping reef bottoms at a coral-reef symposium in Honolulu, Hawaii, on 22 June.

But in the wavelength range applicable to underwater sensing — 430–710 nanometres — Sentinel-2 cannot capture details that CORAL's plane can. The plane carries an instrument that gathers data in more than 100 narrow spectral bands in that range, including the signature

UNDER THE SEA

Over the next three years, a NASA research aeroplane will survey coral reefs throughout the Pacific Ocean — including the rich ecosystems of the Great Barrier Reef in Australia.



of photosynthetic organisms within the living coral itself at 570–575 nanometres.

CORAL will focus on one simple metric: how much coral cover there is on a given reef, as opposed to algae and sand. From that, researchers can calculate how well the coral is doing at transforming sunlight into energy to maintain a reef structure. Hochberg and his colleagues hope to use that information to better understand how local changes, such as an

increase in pollution, might affect coral's health.

The June flights in Hawaii will test whether all the equipment is working. From there, the Gulfstream IV plane will go to the Great Barrier Reef in September and October, followed by Hawaii, the Mariana Islands and Palau in 2017. Divers will simultaneously measure the optical properties of the surrounding seawater and the reef condition up close, to cross-check what the plane sees from 8,500 metres above.

The flights will provide a snapshot of some of the world's most important reefs, says Serge Andréfouët, a marine ecologist at the Research Institute for Development (IRD) in Nouméa, New Caledonia, who led an earlier coral-

mapping effort with the Landsat-7 satellite.

But CORAL will be a one-time glimpse only. With limited funding, there are no plans to repeat any flights to see how the reefs change over time, Hochberg says.

Instead, the team hopes to provide a rich set of baseline data for future coral studies. "You have to pick and choose where you go to try to understand how the ecosystem is working," he says. ■

SOURCE: ERIC HOCHBERG

PUBLISHING

Biology's big funders boost *eLife*

Open-access journal nets £25 million in support until 2022.

BY EWEN CALLAWAY

When three of the world's biggest private biomedical funders launched the journal *eLife* in 2012, they wanted to shake up the way in which scientists published their top papers. The new journal would be unashamedly elitist, competing with biology's traditional 'big three', *Nature*, *Science* and *Cell*, to publish the best work. But unlike these, *eLife* would use working scientists as editors, and it would be open access. And with backers providing £18 million (US\$26 million) over five years, authors wouldn't need to pay anything to publish there.

Four years and more than 1,800 publications later, *eLife*'s funders — the Howard Hughes Medical Institute in Chevy Chase, Maryland, the Wellcome Trust in London and the Max Planck Society in Berlin — announced on 1 June that they will continue their support. They will back the non-profit *eLife* organization

with a further £25 million between 2017 and 2022 (see '*eLife* by the numbers').

"*eLife*'s status in the field is rising quite quickly," says Sjors Scheres, a structural biologist at the Laboratory of Molecular Biology in Cambridge, UK. He became an editor at the journal in 2014, overseeing papers on electron microscopy. "I liked the idea behind it — to make a high-impact journal completely driven by scientists, and open," he says. Although scientists like publishing in the journal, it's less clear whether it has catalysed a wider transformation at the elite end of science publishing.

COLLABORATIVE ATTRACTION

The journal's most innovative feature, according to its authors and reviewers, is its collaborative peer-review process. It turns conventional peer review — in which referees submit individual, and sometimes contradictory, reports — on its head. Instead, referees and scientist-editors work together to identify a submitted paper's

strengths and weaknesses and any needed revisions. Authors receive one decision letter, not individual reports from each referee.

That makes for a speedy review: last year, *eLife*'s published papers took, on average, 116 days from submission to acceptance. For comparison, *Nature* and *Cell* take around 150 days, although *Science* says that in 2013 it took 99 days from submission to acceptance. *Cell* and two of its sister journals have experimented with a similar peer-review model but none has yet adopted it. Peter Binfield, the publisher of another open-access journal, *PeerJ*, in San Francisco, California, says that he likes *eLife*'s peer-review system, but he thinks that the approach would be impossible to scale up to adopt for all published articles.

SELECTIVE BUT OPEN

As it bids to become a top journal, *eLife* has started to turn down more of its submissions. The journal's acceptance rate dropped from

BEN BISHOP

26% in 2014 to 15.4% by 2015, says its editor-in-chief Randy Schekman, a cell biologist at the University of California, Berkeley. That's approaching the acceptance rates of *Nature* and *Science*, which are both below 10%.

In 2013, Schekman denounced *Nature*, *Science* and *Cell* as “luxury journals”, and likened their low acceptance rates and high impact factors to high-end “fashion designers” that artificially stoke demand for their brand through scarcity. Now, he says, *eLife* has become “more selective than I had imagined, but it's not based on any instructions I have conveyed to the editors. It's based on their sensibility of important work.”

In 2014, the most recent year for which financial information is publicly available, *eLife* published 537 research articles with expenses of £3.4 million — equating to around £6,300 for each article. “It appears to be a very expensive way to innovate in the publishing space,” says Binfield.

The journal says that its per-article cost has dropped — to £3,522 in 2015. It points out that it spends money on technology development, too. Six publishers that use the third-party publishing platform HighWire have tested the *eLife*-developed Lens display technology, for instance. Schekman says that *eLife* plans to diversify its income by asking governments and other charities for funding. It will also eventually charge scientists to publish in the journal. But it won't, he says, establish other open-access journals that accept more papers and have lower selectivity — a strategy that some have used to shore up finances. “We have no interest in creating other lesser journals with lower standards,” he says. ■

eLife BY THE NUMBERS

£43 million

Amount committed to the journal over ten years (2012–22) by the Wellcome Trust, Howard Hughes Medical Institute and Max Planck Society.

848

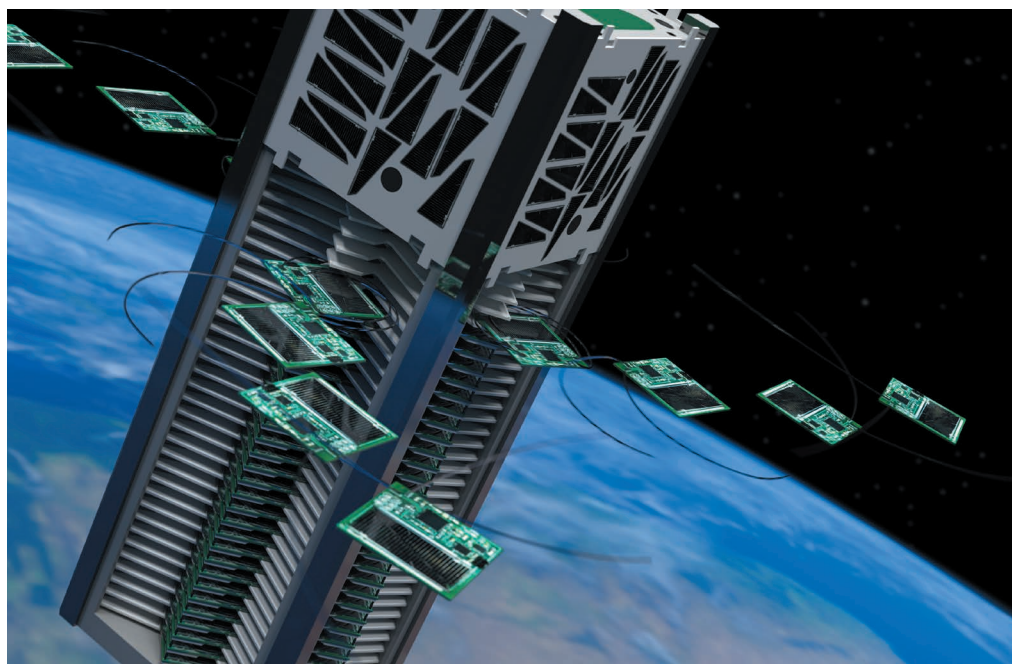
Research articles published in 2015 — all open access.

116 days

Median time to acceptance of paper, 2015.

15.4%

Acceptance rate in 2015.



A KickSat satellite (artist's impression) will launch several minuscule chipsats.

SPACE

First flight for tiny satellites

Launch of ‘chipsat’ probes in July will test a new way to explore the Solar System — and beyond.

BY NICOLA JONES

On 6 July, if all goes to plan, a pack of about 100 sticky-note-sized ‘chipsats’ will be launched up to the International Space Station for a landmark deployment. During a brief few days of testing, the minuscule satellites will transmit data on their energy load and orientation before they drift out of orbit and burn up in Earth's atmosphere.

The chipsats, flat squares that measure just 3.2 centimetres to a side and weigh about 5 grams apiece, were designed for a PhD project. Yet their upcoming test in space is a baby step for the much-publicized Breakthrough Starshot mission, an effort led by billionaire Yuri Milner to send tiny probes on an interstellar voyage.

“We're extremely excited,” says Brett Streetman, an aerospace engineer at the non-profit Charles Stark Draper Laboratory in Cambridge, Massachusetts, who has investigated the feasibility of sending chipsats to Jupiter's moon Europa. “This will give flight heritage to the chipsat platform and prove

to people that they're a real thing with real potential.”

The probes are the most diminutive members of a growing family of small satellites. Since 2003, researchers have launched hundreds of 10-centimetre-sided CubeSats — more than 120 last year alone. Engineer Jekan Thanga at Arizona State University in Tempe is now working on an even smaller ‘femto-satellite’, a 3-centimetre cube that he says has the technological capacity of the first CubeSats. Chipsats, which are smaller and cheaper still, are seen as disposable sensors that could be sent on suicide missions to explore hostile environments, such as Saturn's rings.

“They're all part of the toolbox for next-generation space missions,” says Thanga.

The upcoming chipsat test, called KickSat-2, is the second incarnation of a crowdfunding mission developed by researchers at Cornell University in Ithaca, New York. The shoebox-sized KickSat-1 spacecraft successfully launched on 18 April 2014, but it failed to deploy its cargo of 104 chipsats after a cosmic radiation burst reset the clock on its release mechanism. The ▶

► craft fell out of orbit and burned up with the chipsats still in its hold.

"I was a little bummed out," says Zachary Manchester, an aerospace engineer who built the satellites as a doctoral student in aerospace engineering at Cornell. Fortunately, enough spare parts were lying around to make a second batch relatively quickly and easily.

The chipsats, called Sprites, carry little more than a pair of 60-milliamp solar cells, a radio and an antenna. The KickSat-2 payload includes some newer Sprites that can 'sail' by tilting towards or away from the Sun. A current is run through a coil, turning the chip into a compass needle that aligns with Earth's magnetic field, allowing the chipsat to control its orientation. The probes can be reprogrammed on the fly from the space station.

Sprite prototypes have already proved that they can survive the rigours of space. In 2011, three chipsats were attached to the outside of the space station. They were still working when scientists retrieved them in 2014.

That commercial electronics are good enough to survive space's vacuum and extreme temperatures is a "pretty big deal," says Mason Peck, an aerospace engineer who leads Cornell's chipsat team. But on a flight into deep space, chipsat electronics would face a high risk of damage from radiation. "There are some clear paths to radiation hardening, but it's expensive," says Peck. "And that's not the point. You don't want to make an exquisite satellite. You just launch a million; if only 1% survive then that's fine. You put statistics on your side."

There is plenty of science that Sprites can do closer to home. Peck says that the tiny satellites could be used to verify models of how small bits of debris behave in the upper atmosphere. Like feathers on Earth, the small, flat objects would be heavily affected by drag. "We're not very good at modelling that," says Peck. Another potential project would be to use Sprites to make a high-spatial-resolution map of Earth's magnetic field.

"That would be really useful," agrees Jeffrey Love, a geophysicist with the US Geological Survey in Denver, Colorado, who studies Earth's magnetism. "Ideally you'd want to be measuring it everywhere all the time. This could be a step in that direction."

For the long-term interstellar goal, chipsats will need much better laser-communication capacity. That should be possible, say Peck and Manchester, who are both on the Breakthrough Starshot advisory committee.

"We have gone a long way towards proving we can have a functional tiny craft," says Peck. ■



MATTHIEU COLIN/ITER

The gigantic ITER project is currently under construction in southern France.

NUCLEAR PHYSICS

US urged to stay in fusion project

Department of Energy says US should fund ITER until 2018.

BY DAVIDE CASTELVECCHI
& JEFF TOLLEFSON

The troubled nuclear-fusion experiment ITER has received a cautious vote of confidence from the US Department of Energy (DOE). The multibillion-euro project has improved its performance and management, and the United States should continue to support it, at least until 2018, the DOE said in a report to Congress released on 26 May. But after that, the agency said, the country should re-evaluate its position.

ITER is a collaboration between the European Union, China, India, Japan, South Korea, Russia and the United States. Its goal is to show that fusing hydrogen nuclei to make helium — the same process that heats up the Sun and powers hydrogen bombs — is a technologically feasible way to produce electricity.

The reactor is under construction in St-Paul-lez-Durance in southern France, but the work is more than a decade behind schedule, and its costs have spiralled. The latest report comes against a backdrop of criticism directed at ITER's former management.

The DOE acknowledges ITER's scientific potential, and the substantial improvements

since current director-general Bernard Bigot took over in March 2015. "ITER remains the best candidate today to demonstrate sustained burning plasma, which is a necessary precursor to demonstrating fusion energy power," US energy secretary Ernest Moniz writes in the report's introduction. But the agency says that the progress "must be balanced against several years of inadequate performance". Its recommendation to continue US funding for ITER is contingent on continued and sustained progress on the project, increased transparency and a suite of management reforms.

"I think it's an outstanding report that says all of the right things," says William Madia, a former director of Oak Ridge National Laboratory in Tennessee who led an independent review of ITER in 2013. That report excoriated the way in which ITER was run, and proposed reforms to save it from failure — recommendations that ITER's governing council embraced.

Madia says that the DOE is appropriately encouraged by recent management changes, and appropriately cautious about whether the project is actually back on track. "Bernard is doing a terrific job, but, my goodness, he's got a lot of work to do," he says. Bigot acknowledges this, and says that the DOE's conclusions are

the most he could have hoped for at this point: “We know there is still a long way to go.”

The DOE is a major funder of fusion research. But although the United States is bound by an international treaty to provide its share of ITER's costs — a relatively small 9% of the project's budget — it cannot meet its contributions if Congress does not approve them.

GROWING BUDGET

The report's recommendations have provoked scepticism on Capitol Hill. Senator Dianne Feinstein of California, the highest-ranking Democrat on the Senate panel that oversees DOE spending, says that the United States cannot afford to keep pace with ITER's growing budget. The DOE estimates that the country's annual contribution, currently US\$115 million, will more than double by 2018.

Last year, the Senate proposed to end support for ITER, but backed down during final negotiations with the House of Representatives. This year, it is not clear that ITER will win a reprieve. On 12 May, the Senate approved an energy-funding bill for fiscal year 2017 that cut all spending on ITER. And on 26 May, the House rejected its own 2017 energy-spending bill, which included money for ITER.

Without the United States, ITER would probably survive, says Mark Koepeke, a plasma physicist at West Virginia University in

Morgantown who leads a government advisory panel on fusion research. But in April, Bigot told US lawmakers that the country's fusion expertise would be difficult to replace. Madia says that the effect of a US exit is impossible to predict: “It makes good cocktail conversation, but no one knows what would actually happen.”

“ITER remains the best candidate today to demonstrate sustained burning plasma.”

ITER's approach to fusion is to trap heavy isotopes of hydrogen in a doughnut-shaped vacuum vessel called a tokamak and heat them to 150 million °C. This should force their nuclei to fuse, releasing vast amounts of energy. Other tokamaks exist, but ITER would be the first to release substantially more energy than was put into the hydrogen plasma.

Begun in 2007, the project was originally due to be completed in 10 years for €5 billion (US\$5.6 billion). Observers say that under previous director-general Osamu Motojima, who was in office from 2010 to 2015, the experiment was in denial about slipping deadlines and witnessed a drop in staff morale. After the independent review by Madia, the ITER Council accelerated the transition to a new director-general, nominating Bigot, a French

nuclear physicist with extensive management experience, in late 2014.

By November 2015, Bigot's team had presented a revised timetable for the project, and estimated that it would cost an extra €4.6 billion to bring to completion. The team said that the earliest possible date for getting hydrogen plasma to run inside the machine was 2025, and that it would take several more years to inject the heavier hydrogen isotopes tritium and deuterium, and achieve fusion.

In April, an external review from the ITER Council Working Group confirmed that progress had been made on the recommendations of the Madia report, and that the new management had been realistic about the earliest possible date for plasma. But it pointed out that the estimates of costs and the completion date did not take into account possible contingencies.

The latest DOE report recommends funding the cost increases cited by Bigot, but remains sceptical about the schedule. It outlines two funding scenarios: one based on achieving first plasma in 2025, and a more realistic scenario that pushes the date back to 2028.

Bigot's team also proposed a more modest plan, which achieves first plasma on time but delays fusion. This should save money by postponing the parts of construction that are not needed for first plasma, but no one has yet calculated how much. ■

MATHEMATICS

Maths proof smashes size record

Supercomputer produces a 200-terabyte proof — but is it really mathematics?

BY EVELYN LAMB

Three computer scientists have announced the largest-ever mathematical proof: a file that comes in at a whopping 200 terabytes, equivalent to all the digitized text held by the US Library of Congress. The researchers have created¹ a 68-gigabyte compressed version of their solution — which would allow anyone with about 30,000 hours of spare processor time to download, reconstruct and verify it — but a human could never hope to read through it.

Computer-assisted proofs too large to be directly verifiable by humans have become common, as have computers that solve problems in combinatorics — the study of finite discrete structures — by checking through umpteen individual cases. Still, “200 terabytes is unbelievable”, says Ronald Graham, a mathematician at the University of California, San Diego. The previous record-holder is thought to be a 13-gigabyte proof², published in 2014.

The puzzle that required the 200-terabyte proof, called the Boolean Pythagorean triples

problem, has troubled mathematicians for decades. In the 1980s, Graham offered a prize of US\$100 for anyone who could solve it. (He presented the cheque to one of the three computer scientists, Marijn Heule of the University of Texas at Austin, last month.) The problem asks whether it is possible to colour each positive integer either red or blue, so that no trio of integers a , b and c that satisfy Pythagoras' famous equation $a^2 + b^2 = c^2$ are all the same colour. For example, for the Pythagorean triple 3, 4 and 5, if 3 and 5 were blue, 4 would have to be red. ►



VIDEO



Neanderthals built cave structures — and no one knows why go.nature.com/caveq

MORE NEWS

- Why the deal to fix US chemical laws matters go.nature.com/chemregs
- Cloud-seeding surprise could improve climate predictions go.nature.com/clouds
- Jailed Iranian physicist released on bail go.nature.com/kokabee

NATURE PODCAST



The genetics of moth evolution, Earth's core conundrum and Pluto's polygonal surface nature.com/nature/podcast

► In a paper¹ posted on the arXiv server on 3 May, Heule, Oliver Kullmann of Swansea University, UK, and Victor Marek of the University of Kentucky in Lexington show that there are many allowable ways to colour the integers up to 7,824 — but when you reach 7,825 or above, it is impossible for every Pythagorean triple to be multicoloured. There are more than $10^{2,300}$ ways to colour the integers up to 7,825, but the researchers took advantage of symmetries and several techniques from number theory to reduce the number of possibilities that the computer had to check to just under 1 trillion. It took about 2 days running 800 processors in parallel on the University of Texas's Stampede supercomputer to zip through all the possibilities. The researchers then verified the proof using another computer program.

The Pythagorean triples problem is one of many similar questions in Ramsey theory, an area of mathematics that is concerned with finding structures that must appear in sufficiently large sets. For example, the researchers think that if the problem had allowed three colours, rather than two, they would still have hit a point where it would have been impossible to avoid creating a Pythagorean triple where a , b and c were all the same colour; indeed, they conjecture that this is the case for any finite choice of colours. Any proof for more colours will probably be even larger than the 200-terabyte 2-colour proof, unless researchers can simplify the case-by-case checking process with a breakthrough in understanding.

Although the computer solution has cracked the Boolean Pythagorean triples problem, it hasn't provided an underlying reason why the colouring is impossible, or explored whether the number 7,825 is meaningful, says Kullmann. That echoes a common philosophical objection to the value of computer-assisted proofs: they may be correct, but are they really mathematics? If mathematicians' work is considered to be a quest to increase human understanding of mathematics, rather than to accumulate an ever-larger collection of facts, a solution that rests on theory seems superior to a computer ticking off possibilities.

In the case of the 13-gigabyte proof² from 2014, which solved a special case of a question called the Erdős discrepancy problem, a theory-based solution was eventually found. Mathematician Terence Tao of the University of California, Los Angeles, solved³ the general problem the old-fashioned way in 2015 — a much more satisfying resolution. ■

1. Heule, M. J. H., Kullmann, O. & Marek, V. W. Preprint at <http://arxiv.org/abs/1605.00723> (2016).
2. Konev, B. & Lisitsa, A. Preprint at <http://arxiv.org/abs/1402.2184> (2014).
3. Tao, T. Preprint at <http://arxiv.org/abs/1509.05363> (2015).



Napthalene is one of the chemicals slated for review by the US Environmental Protection Agency.

EXPLAINER

US chemicals law set for overhaul

Bill would give government more authority to regulate potentially toxic substances.

BY JEFF TOLLEFSON

The US Congress is poised to overhaul the law that governs the introduction and use of chemicals, in one of the most significant changes to the country's environmental regulation in decades. The update to the 1976 Toxic Substances Control Act (TSCA) comes after more than ten years of debate, and many failed attempts to revamp the law.

The US House of Representatives passed the bill with overwhelming bipartisan support on 24 May. The Senate is expected to approve the measure soon, clearing the way for US President Barack Obama to sign it into law.

Nature takes a look at the implications of the historic deal, which will give the US Environmental Protection Agency (EPA) new power to ensure that chemicals — both old and new — are safe.

Why amend the current law?

Critics of the TSCA have long complained that the law effectively ties the EPA's hands, preventing the agency from examining the safety of known chemicals and making it difficult to ensure that new ones do not pose undue health hazards.

The law requires companies to register new

chemicals before they are used in products and industrial processes, but the default assumption is that all chemicals are safe. Unless the EPA can show that a given chemical poses an unreasonable risk to human health or the environment, that chemical is automatically approved for use. Companies do not have to provide the agency with much information about their chemicals, and the EPA cannot require industry to conduct additional research without solid evidence that a chemical poses a health risk.

How many chemicals does the EPA regulate?


Companies introduce about 700 chemicals into the marketplace each year. And 40 years after the TSCA became law, the EPA's chemical inventory lists 85,000 substances. But nobody knows exactly how many of these chemicals are still in use.

The EPA has identified 90 chemicals that merit further investigation, and possibly regulation. But only about 2% of the chemicals in use today have undergone a safety review by government scientists, according to the Environmental Defense Fund, a watchdog group in New York City.

So, what will change?

In short, everything. Once the TSCA is

KEITH WHEELER/SPL



amended, the EPA will have the authority to ask questions, seek more information and even require companies to conduct additional studies to ensure that chemicals are safe.

The US law requires less information about chemicals up front than Europe's pioneering chemical-safety legislation, but the two regulatory approaches should have similar results, says Richard Denison, a biochemist who tracks chemical safety for the Environmental Defense Fund.

"The EPA now has to make an affirmative finding that a chemical is safe in order for that chemical to go on the market," he says. Moreover, Denison notes, the legislation allows the EPA to determine the risks posed by a chemical without considering the economic implications of that decision.

The revised law will also restrict companies' ability to withhold information about chemicals from the public by arguing that the data are a trade secret. Whereas most claims for confidentiality sail through under the current law, Denison says that the amended statute will require firms to provide detailed explanations for why the information they submit to the EPA should remain secret.

This change, along with the EPA's new ability to review more chemicals, will give researchers and the public greater access to information about chemicals in the environment.

How did the new bill arise?

Lawmakers in Congress have debated whether — and how — to revise the TSCA since at least 2005. But repeated efforts to overhaul the law failed as politicians debated how to expand the EPA's authority to regulate chemicals without stifling commercial innovation.

The chemical industry initially opposed efforts to reform the TSCA, but gradually changed its position as state and international chemical regulation expanded. When a Republican proposal to amend the TSCA gained momentum in 2013, Democrats began to join the effort and a compromise slowly emerged.

The House passed its own TSCA reform bill in June, and the Senate approved its bill legislation in December. For months, lawmakers have been hammering out a compromise measure that blends aspects of the House bill with the major components of the Senate legislation.

The result is a deal with broad bipartisan backing in Congress, and endorsements from the White House and industry. Many environmental groups have expressed cautious support for the legislation, but remain concerned about its recommended funding levels, the continuing requirement to consider costs when developing regulations, and provisions that could allow the federal government

to override state regulation of chemicals.

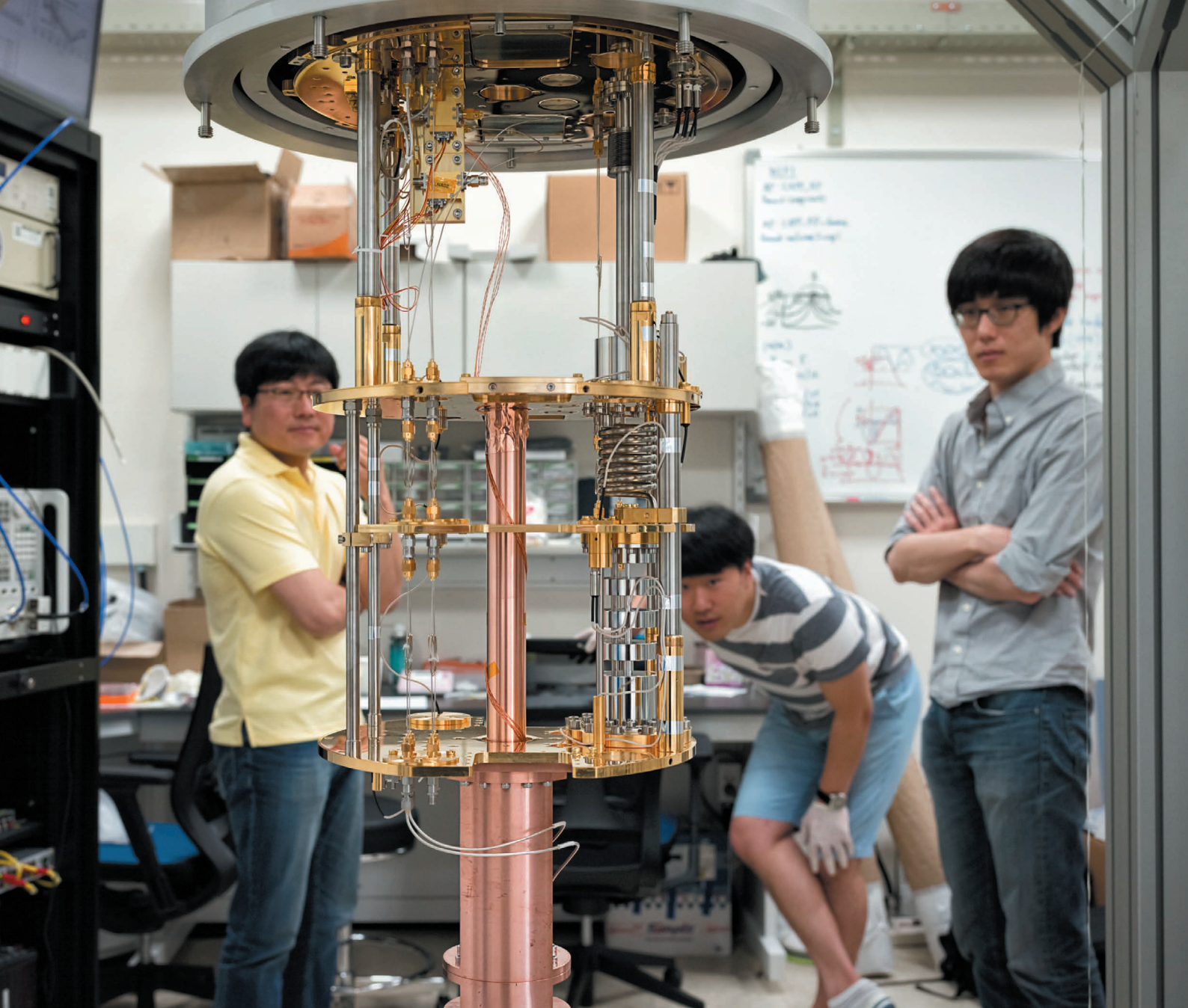
What comes next?

After the bill is enacted, the EPA will draw up rules for its new review process, which includes determining the fees that companies will need to pay to submit chemicals for government review. The legislation allows the agency to collect up to US\$25 million per year in fees to supplement its budget for chemical regulation, which is intended to cover roughly one-quarter of the total programme cost.

The EPA must also figure out which of the 85,000-odd chemicals in its inventory are still in use. The agency will survey companies that make and use chemicals to revise its list. Once that's done, agency scientists can go through the list and prioritize those chemicals that merit a safety review. ■

CORRECTION

The figure given for the planting of super soya bean in the News Feature 'Frugal farming' (*Nature* **533**, 308–310; 2016) should have been 67,000 hectares, not 1 million. In addition, the feature failed to make it clear that Jonathan Lynch was joking when he suggested that students should "drop acid".



South Korea's Nobel dream

The Asian nation spends more of its economic output on research than anywhere else in the world. But it will need more than cash to realize its ambitions.

BY MARK ZASTROW

Behind the doors of a drab brick building in Daejeon, South Korea, a major experiment is slowly taking shape. Much of the first-floor lab space is under construction, and one glass door, taped shut, leads directly to a pit in the ground. But at the end of the hall, in a pristine lab, sits a gleaming cylindrical apparatus of copper and gold. It's a prototype of a device that might one day answer a major mystery about the Universe by detecting a particle called the axion — a possible component of dark matter.

If it succeeds, this apparatus has the potential to rewrite physics and win its designers a Nobel prize. "It will transform Korea, there's no question about it," says physicist Yannis Semertzidis, who leads the US\$7.6-million-per-year centre at South Korea's premier technical university, KAIST. But there's a catch: no one knows whether axions even exist. It's the kind of high-risk, high-reward project

SHIN WOONG-JAE

A prototype axion detector in Daejeon, South Korea. that symbolizes the country's ambition to become a world leader in basic research.

South Korea is spending heavily to achieve its goal. In 1999, the country's investment in research and development (R&D) totalled 2.07% of its gross domestic product (GDP), just below the average for nations in the Organisation for Economic Co-operation and Development (OECD). In the latest figures, the country has stretched out a clear lead at the top. The 4.29% (63.7 trillion won, or US\$60.5 billion) that South Korea invested in R&D in 2014 outstrips runner-up Israel (at 4.11%), as well as regional competitor Japan and the United States. The biggest chunk of the money goes towards applied research and development in industry, but the government has made major investments in basic science, too.

The big hope is that the country can innovate its way out of a looming economic crisis — and win a Nobel prize in the process. South Korea aims to increase its investment to 5% of GDP by 2017, and last month, President Park Geun-hye's government announced that it would boost annual basic-science funding levels by 36% by 2018, to 1.5 trillion won. “Basic research starts with intellectual curiosity among scientists and technicians, but it could be a source of new technologies and industries,” Park said.

Can the country achieve its ambition? That depends who you ask. Some Korean scientists and policymakers doubt that it can sustain its high level of investment, and they worry that cultural barriers and bureaucracy are hindering research. Young scientists are voting with their feet: according to figures released in 2014 by the US National Science Foundation (NSF), nearly 70% of South Koreans who were awarded PhDs in the United States in 2008–11 planned to stay there.

Reorienting the nation's science focus is no easy task, says Youngah Park, president of the Korea Institute of S&T Evaluation and Planning (KISTEP), a government think tank in Seoul. The country has long been an industry-focused ‘fast follower’ — excelling at quickly adopting technologies and products, such as semiconductors and smartphones, and making them better and cheaper. Now, Korea needs a new model, she says. “That is a very challenging and adventurous scheme for us.”

SHOCK AND AWE

When the artificial-intelligence (AI) program AlphaGo beat Korean grandmaster Lee Sedol at the game Go this March, the impact on the national psyche was profound. The AlphaGo shock, as it came to be known, showed the country that AI was the future: Korea must catch up to the likes of Google DeepMind in London, which invented the Go-playing machine.

Within days, President Park announced that the government would invest 1 trillion won in

AI by 2020, and prod the private sector into investing a further 2.5 trillion won. The initiative's cornerstone would be a public–private research institute involving corporations such as Samsung and LG. But many scientists criticized the approach as a knee-jerk reaction that would funnel government money into product development, not into the type of basic research that the country needs.

The funding injection was typical of the strategy that has propelled South Korea's economy over the past few decades: the government set goals and then channelled money to corporate partners to carry them out. The formula was devised by Park's father, dictator Park Chung-hee, who seized power in a 1961

“We have large funding and you can do what ever you want to.”

coup. During his 18-year reign, he favoured companies that grew into behemoths — conglomerates, called *chaebol* in Korea, such as Samsung, LG and Hyundai, which remain the backbone of the nation's economy today.

Powered by these industries, five decades of economic growth vaulted South Korea from developing-world poverty to membership of the group of 20 (G20) leading industrial nations. As the country moved painfully from dictatorship to democracy, government support for research remained a bipartisan priority — mainly as a driver for further growth. Korea's corporate giants still dominate the R&D scene. According to KISTEP figures, of the 63.7 billion won spent on R&D in 2014, 49.2 billion came from private enterprises. That includes more than half of the 11.2 billion won spent on basic research. Much industrial research happens behind closed doors, although partnerships with academia are on the rise.

Meanwhile, government-funded labs also worked mainly towards developing industrial technologies, and blue-sky, basic research remained an afterthought. “Politicians don't distinguish between R&D in technology and support in basic science,” says physicist Doochul Kim. Until recently, he says, “there has been no support in basic science, basically”.

Change arrived during the run-up to the 2007 presidential election, when a group of researchers pitched an idea to the nation's leading politicians: that the country build an Institute for Basic Science (IBS). The organization would be Korea's answer to Germany's academically elite Max Planck institutes and Japan's RIKEN centres. “It was the first time that scientists went forward and suggested their

own big project for the nation,” says Youngah Park, who was a legislator with the conservative party at the time. The institutes would be part of an even bigger plan to create a research and business megahub called the International Science and Business Belt — and this became government policy when conservative candidate Lee Myung-bak won the election.

Political wrangling subsequently forced the government to scale back some of its plans, but IBS survived, in modified form. Fifty IBS centres, one-third of them in Daejeon, would be funded at an average of 10 billion won a year each for at least 10 years — a boon for researchers, who would be offered secure support to pursue their ideas. “We have large funding and you can do whatever you want to,” says Kim, now president of the IBS. Today, 26 of the centres have opened, with the rest hoped to follow by 2021.

AXION RACE

The Center for Axion and Precision Physics (CAPP) at KAIST is one of them. Semertzidis became head of the centre in 2013, moving from Brookhaven National Laboratory (BNL) in Upton, New York.

In its quest to find the axion, CAPP is chasing a high-profile rival in the United States: the Axion Dark Matter Experiment (ADMX), based at the University of Washington in Seattle. “Very smart people, absolutely,” Semertzidis says, with a disarming grin. “But we'll win, nonetheless — absolutely.”

If axions are indeed part of dark matter, they should be all around us. CAPP's design — like that of the ADMX — uses a cavity that should resonate at the axion's mass, with strong magnets outside it that cause the particles to convert into two photons and pop into sight. But physicists don't know what the axion's mass is, so they have to scan for it, tuning the resonant frequency of the cavity with rods of copper or sapphire. It will take years for a single device to cover the whole range of possible frequencies.

CAPP has at least a year of development left, whereas ADMX is already beginning operations, giving it a significant head start. But CAPP plans to build not one, but seven cavities — all in that hole in the ground, down the hall. And it has more powerful magnets, developed at BNL. “We'll do it seven times better, because of the sheer power of money,” says Semertzidis, who thinks that his team can leapfrog the ADMX within five years. ADMX leader Leslie Rosenberg suspects that it could, too. “CAPP is by far our most credible competitor,” he says.

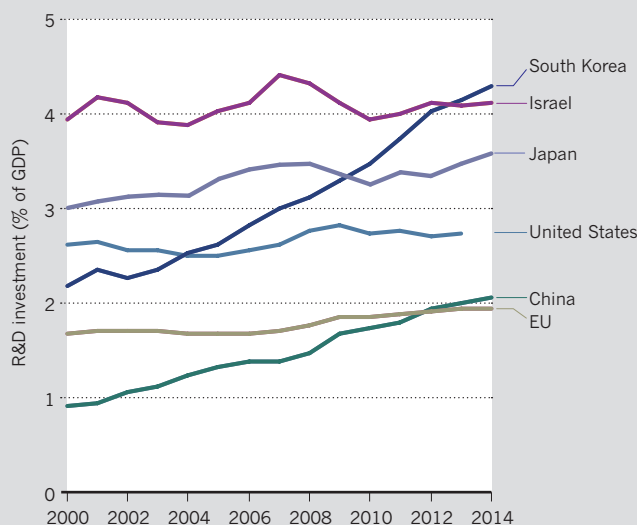
Whatever the outcome, Rosenberg says that CAPP's progress is a milestone for Korean physics. The country's willingness to spend landed foreign talent and technology. “These new IBS centres have moved them into the top tier,” he says. The other 25 existing centres are pushing into fields ranging from gene editing to nanomaterials and pure mathematics. Roughly one-third of the IBS's budget

Science in South Korea

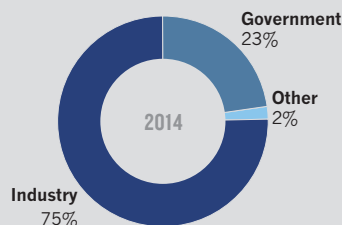
Industrial research and development (R&D) has long been a priority for South Korea as a driver of economic growth. In the past decade or so, more emphasis has been placed on basic research.

R&D investment

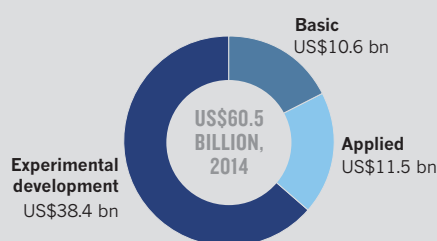
South Korea's spending on R&D has soared to more than 4% of its gross domestic product (GDP) — more than any other country in the world and double that of China and the European Union.



Most R&D money comes from industry ...

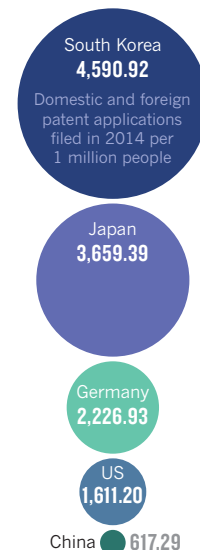


... and more goes to applied than to basic research.



Patents

South Korea is a world leader in patent applications, in part thanks to industry leaders such as Samsung and LG.



is devoted to one flagship effort — the Rare Isotope Science Project (RISP) in Daejeon, which seeks to build a heavy-ion accelerator for nuclear science and biomedical research.

South Korea is also investing in basic-research facilities outside of IBS. The Pohang Accelerator Laboratory is receiving a 400-billion-won upgrade to house an X-ray free-electron laser that can image materials on nanometre and femtometre scales. And in 2014, the nation completed construction on a 107-billion-won, state-of-the-art Antarctic research centre in Terra Nova Bay, which quickly became the envy of the polar-research community. “It was like a spaceship had landed,” said US National Science Foundation polar-science head Kelly Falkner in 2014, months after attending the sleek facility’s opening ceremony. “It’s amazing to see what they can do by starting from scratch.”

Rosenberg, for one, says that Korea is wise to invest in the IBS centres. “If they can continue to afford it, I think the pay-off is going to be enormous.” And if they find the axion? “Oh my goodness, well, let’s say it would instantly be a Nobel prize.”

And that is something that this country wants very much indeed.

NOBEL DREAMS

Last October’s Nobel-prize announcements triggered a wave of disappointment — again. There were no awards for South Korean researchers, but scientists in Japan, the nation’s most bitter regional rival, collected shares in two: Satoshi Ōmura for developing

a therapy for roundworm, and Takaaki Kajita for showing that neutrinos have mass. “Why no Korean Nobel laureates?” asked a headline in *The Korea Times*.

The question came up again at an oversight hearing of South Korea’s parliamentary science committee, held that week. One member of parliament compared the full list of the two countries’ Nobel laureates in science to a dismal football result: Japan 21, South Korea 0. “When will IBS score a goal?” he asked Kim.

“Oh my goodness, well, let’s say it would instantly be a Nobel prize.”

In some political quarters, IBS was originally hailed as a way to level the Nobel score, but Kim has pushed back against that, arguing that the ‘Nobel complex’ leads to shortsighted policies that chase hot topics and demand instant results. “We are only four years old,” he told the committee. He noted that it took decades to develop the infrastructure at Japan’s Kamiokande Observatory near Hida, where the neutrino breakthrough was made. “So you shouldn’t ask that question,” he said.

Korea did seem poised for a Nobel just over a decade ago, when stem-cell scientist

Woo Suk Hwang claimed to have derived the world’s first stem-cell lines from cloned human embryos. But glory quickly turned to shame when Hwang was first found guilty of ethics violations in the way he collected women’s eggs for research, and then discovered to have fabricated some of his work. The scandal left the impression that the country’s oversight of research ethics and integrity was lax.

Scientists in Korea say that the scandal has brought about positive changes. Slowly, more Korean journals have begun to issue retractions, says Eric di Luccio, a structural biologist at Kyungpook National University in Daegu, and many universities are using the plagiarism-detection site turnitin.com to check papers and theses. More attention is also being paid to bioethics, says Jin-Soo Kim, director of the IBS Center for Genome Engineering at Seoul National University. “Before the Hwang scandal, in the laboratory, people would just draw blood and do experiments,” he says. “Now it’s recognized that you shouldn’t do it without approval” from an institutional review board.

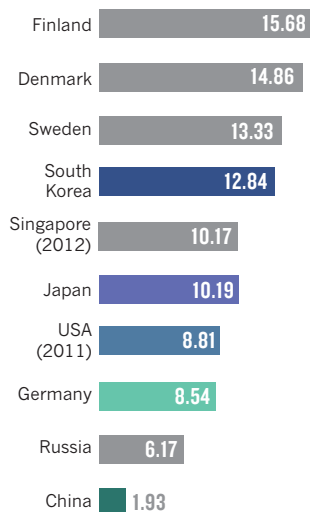
But Kim says that one ‘Hwang-gate’ reform is now holding Korea back: in the wake of the scandal, the government enacted a ban on human-embryo research, with only occasional exceptions granted for stem-cell studies. Kim has been at the forefront of developments in CRISPR-Cas9 gene editing, a technique that is revolutionizing biomedical research, but he has found himself unable to use the technology for research in human embryos, even as teams in China, the United Kingdom and elsewhere forge ahead with such work. “It’s a pity,” says

SOURCE: R&D INVESTMENT: OECD; APPLIED/BASIC RESEARCH: KISTEP; WORKFORCE: UNESCO; PUBLICATIONS: SCOPUS

Workforce

South Korea has one of the world's highest proportions of researchers.

Researchers per thousand people in employment, 2013



Kim, who has instead focused his efforts on engineering pigs and plants.

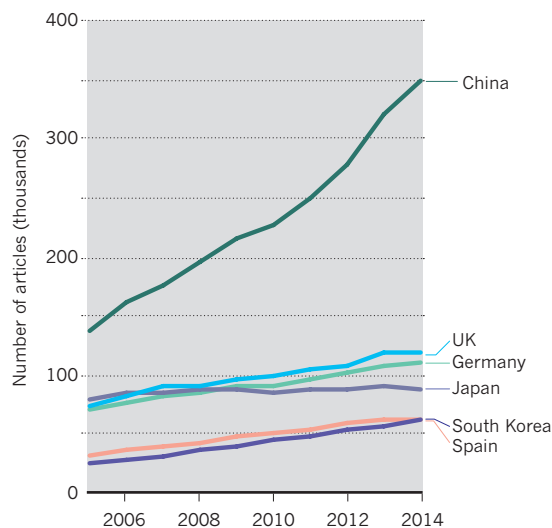
Researchers also chafe at other regulations. At public universities, tenure and promotion decisions are often based in part on evaluations that count papers by fractional contribution: a four-author paper, for example, would earn a scientist a small fraction of the credit of a single-author one. The system is “rather counterproductive”, says di Luccio. It dissuades scientists from taking part in the large international collaborations of modern big-budget science, and encourages them to publish single-author papers in less-prestigious national journals to juice up their evaluation scores. “I did it three times already,” he says. “This evaluation system is the exact opposite of what it should be to elevate scientific research.” The government says that universities and other organizations are free to implement their own standards of evaluation and that nationally funded programmes use more qualitative measures. (IBS insulates researchers from paper counts.)

Some scientists see deeper problems with the academic culture, rooted in Korean society at large. Secondary and undergraduate education focus on test-taking and emphasize deference to teachers — tendencies that academics bemoan as discouraging the creativity and debate necessary in a lab. “When new students come, they are quiet — that is the Korean culture,” says Jin-Soo Kim, who counters this by requiring his students to ask questions before they can leave group meetings.

Korean customs were a turn off for Young-Im Kim, who was doing a physics postdoc at

Publications

South Korea has more than doubled its academic publication output since 2005, overtaking similarly populated Spain — but lagging behind its regional rival Japan. Scientists publish most in chemistry, engineering, physics and life sciences.



the University of Oxford, UK, in 2014, when a friend sent her a link to a job posting at CAPP. Although she thrilled to the research, she was hesitant to return to her home country because of the hierarchical nature of Korean culture. “The only reason I applied is because of Yannis,” she says. “If he were Korean, I wouldn’t have.” She is now a research fellow at CAPP.

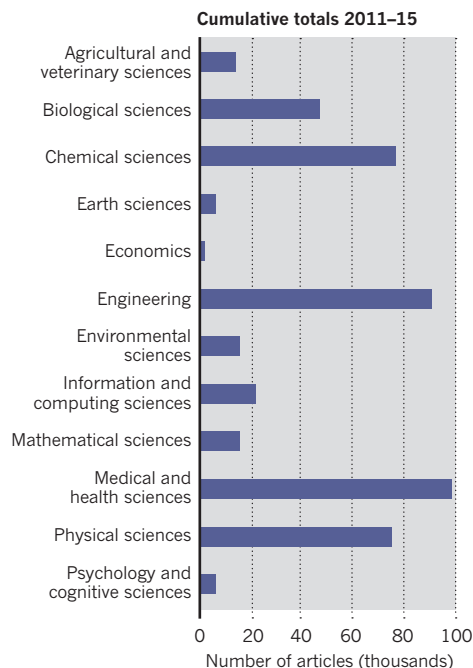
Cultural barriers can have a disproportionate impact on female scientists. One example, says Young-Im Kim, is Korean drinking culture, in which men often stay out late with their male co-workers. Important workplace decisions are often made at such events, effectively excluding women. Such problems could go some way towards explaining why Korea has a wide gender gap in its scientific workforce. According to OECD figures, in 2010 less than 17% of researchers in South Korea were women. In Portugal, the OECD leader, the fraction is 45.5%.

STRETCHED RESOURCES

Policy analysts warn that research spending may slow in the future, as Korea faces the likely prospect of an economic slowdown and, in the long term, a social-welfare net stretched to support an ageing population with one of the lowest birthrates in the world. And although R&D expenditure continues to grow as a percentage of GDP (see ‘Science in South Korea’), it is actually shrinking when viewed as a percentage of government spending, says Youngah Park. “That is a sign that we have no room to increase this government R&D budget anymore.”

Some critics say that spending is now too

Publications by discipline



focused on IBS: that the bold plan is sucking up basic-research funds and reducing the pot of money available through other grants, creating a situation of haves and have-nots, and potentially quenching original — perhaps Nobel-worthy — projects. The shrinking grant pool is a legitimate issue, acknowledges Doochul Kim. But he thinks that the government should address it by shifting funds from applied research. It continues to subsidize such research when many say that it should be focusing on long-term basic research that industry wouldn’t pursue.

“There is some excellent science done in Korea, but still, in general, the average is not as good as the advanced countries like the US, UK and Germany,” says Jinwoo Cheon, director of the IBS Center for Nanomedicine at Yonsei University in Seoul. To spur investment in basic research, he adds, scientists have to convince the public and government officials of its intangible benefits. “Excellence in basic science is not easy to have, and it has to be rooted in our society — curiosity-driven research, and knowing different ways of thinking.”

Sunchan Jeong, director of RISP, says that if there is such a thing as a recipe for winning a Nobel prize, then IBS has got it. “Select some competitive fields in the world and concentrate their investment on it. That’s a good way.” But there are no guarantees, he cautions: “The people in Korea should understand that scientific results are not necessarily repaid by some greater prize like the Nobel.” ■

Mark Zastrow is a writer based in Seoul.



After a sneeze, large droplets of saliva and mucus shoot out of the mouth, but fall relatively quickly.

A turbulent cloud carries smaller droplets and allows them to drift for up to 8 metres.

Mathematician *Lydia Bourouiba* uses high-speed video to break down the anatomy of sneezes and coughs, and to explore how diseases spread.

WHERE SNEEZES GO

BY CORIE LOK

So, how do you get your research subjects to sneeze on cue? “That’s a question I get a lot,” says Lydia Bourouiba with an easy smile. The solution turns out to be surprisingly simple: just take a small, rod-shaped device, use it to tickle a subject’s nostril for a few seconds, and — achool!

For Bourouiba, a mathematician and fluid dynamicist, that sneeze is the pay-off. She and her team at the Massachusetts Institute of Technology (MIT) in Cambridge record the explosive aftermath in gross detail using one or sometimes two cameras running at thousands of frames per second. Played back in slow motion, the videos reveal a violent explosion of saliva and mucus spewing out of the mouth in sheets that break up into droplets, all suspended in a turbulent cloud.

The videos that Bourouiba has recorded in this way allow her to

measure everything from the diameter of the droplets to their speed — data that help her to learn more about how these particles carry viruses and other pathogens to their next host. She has shown that sneeze and cough particles can travel the length of most rooms and can even move upwards into ventilation shafts — suggesting that microbes in the droplets could potentially spread farther and over longer periods of time than current theories suggest.

Ultimately, says Bourouiba, her goal with this work is to ground epidemiology and public health in physics and mathematics. When trying to keep diseases from running rampant, she says, “we want to be giving recommendations that are based on science that has been tested in the lab”. In practical terms, such insights could lead to maps showing the contamination risks in the vicinity of infected people, protective equipment optimized to shield hospital workers from specific

A sneeze captured on high-speed video.

kinds of germs, and better predictions of how diseases move through a population.

Bourouiba pursues this goal with the same energy and ambition that leads her to fill her leisure time with week-long bike trips, mountain climbing — she ascended Tanzania's Mount Kilimanjaro in 2011 — and winter camping at -20°C . Although she is hardly the first researcher to use high-speed video to study fluid dynamics, she is the first to realize its potential in the respiratory field, says David Ku, a biofluid-mechanics researcher at Georgia Institute of Technology in Atlanta. Bourouiba's approach could be transformative in the field, says Ron Fouchier, a virologist at the Erasmus University Medical Center in Rotterdam, the Netherlands. "This kind of physics is absolutely needed to understand how transmission works."

A FLUID CAREER

Bourouiba has been a natural explorer as far back as she can remember. As a child in France, she immersed herself in books about science and nature, including a biography of Albert Einstein. She soon fell in love with mathematics and physics, and made them her major subjects when she earned her undergraduate degree in France and Montreal, Canada.

But during her graduate work in fluid mechanics at Montreal's McGill University, as she focused on narrower and narrower theoretical questions about turbulent flows, Bourouiba began to feel the itch for something more. She had spent some of her early years in Algeria during the civil war of the 1990s, and vividly remembered the turmoil and misery that she had witnessed there. "We know what the worst is, we saw a lot of it," she says. "But what can we as a species do to push that boundary of what we can achieve, in terms of making the world a better place?"

In her search for an answer, Bourouiba soon homed in on health and epidemiology. This was the mid-2000s, and emerging diseases were all over the news. Severe acute respiratory syndrome (SARS) had killed nearly 800 people around the world in 2003, polio was making a comeback and avian flu was jumping across to humans. Infectious diseases seemed to Bourouiba like the perfect way to combine all of her interests and expertise.

She was tentative at first. A career in fluid mechanics promised to be secure and certain, whereas a head-first dive into biology seemed like a huge risk. But one day, about halfway through her PhD, she was mulling this conundrum as she made her way up the wall at a rock-climbing gym. "So what?" Bourouiba suddenly said to herself as she reached for the next handhold. "You can't make decisions out of fear."

VIOLENT EVENTS

Having come so far, Bourouiba saw her fluid-dynamics PhD to completion in 2008. But from there she managed to land a postdoc appointment in mathematical epidemiology at York University in Toronto, Canada, where she started thinking about sneezes and coughs.

These 'violent expiratory events' (as one of Bourouiba's papers calls them) were assumed to be one of the main ways that respiratory diseases spread. But how, exactly? Epidemiological studies estimate how a disease is transmitted on the basis of people's movements and activities at the time they got infected. Did they contract the disease by direct, person-to-person contact, such as shaking someone's germ-covered hand, or from contaminated surfaces such as doorknobs? Was it through large droplets that make a short leap from one respiratory tract to another, or through smaller aerosol particles that are suspended in air and can travel farther before being inhaled? Or was the route some combination of these modes?

Such studies have helped researchers to work out that measles is typically spread by aerosols and that Ebola is transmitted mainly through direct contact with infected bodily fluids. But there is still a lot of uncertainty for many pathogens, which hampers the ability of public-health officials to control the spread of disease during outbreaks and to prepare for future ones. SARS, for example, is thought to spread mainly

through close contact, yet the 2003 outbreak showed at least some evidence of airborne transmission¹. And some researchers think that Ebola viruses might travel through air to some degree².

At York, Bourouiba became convinced that these uncertainties could be reduced by pinning down some key details about the physics of sneezes and coughs that conventional disease-transmission models are missing.

In 2010, a postdoc appointment at MIT gave her a chance to start filling those gaps with hard data. Up to that point, she had worked only on theory — but now she plunged into experimental research, learning through trial and error the subtleties of high-speed video and lighting to capture a sneeze. "Mathematicians are often uncomfortable in a lab setting," says John Bush, a fluid dynamicist who was her mentor at MIT. "Lydia really took to it."

SUSPENDED SPRAY

One thing that Bourouiba particularly wanted to pin down was the size distribution of the droplets coming out of the mouth, because size affects how many microbes a droplet can carry and how far it can travel through the air.

For her first set of experiments, published in 2014, she wanted to look at the entire spray of droplets³. Bourouiba posted adverts around the MIT campus to recruit volunteers, and filmed the coughs and sneezes of about ten healthy people. After much tinkering with camera positions, backgrounds and lighting levels — at one point, the lights made the room uncomfortably hot for participants — Bourouiba captured videos that showed that the droplets were propelled out of the mouth in a turbulent, buoyant cloud. The cloud grew and slowed down as it pulled in air from the environment, lifting and carrying the droplets away from the sneezer.

The video evidence contradicted conventional thinking about sneezes, which held that larger droplets would fall to the ground within 1–2 metres, and that only the smaller ones would stay aloft as airborne aerosols. Feeding her video evidence into her mathematical models, Bourouiba concluded that, thanks to the cloud dynamics, many of the larger droplets can travel up to 8 metres for a sneeze and 6 metres for a cough, depending on the environmental conditions, and stay suspended for up to 10 minutes — far enough and long enough to reach someone at the other end of a large room, not to mention the ceiling ventilation system.

That conclusion has implications for health-care workers, says James Hughes, an infectious-disease epidemiologist at Emory University in Atlanta. If a disease is thought to be transmitted within 1–2 metres, workers might assume that they are safe beyond that zone. "I think maybe we need to be a little bit more circumspect about that," he says.

For Bourouiba's next set of experiments⁴, she zoomed in closer to the mouth to film a 150-millisecond-long sneeze. Videos taken from the side and top at up to 8,000 frames per second revealed that the fluid breaks up in steps, like a slow-motion explosion produced by Hollywood: the fluid emerges from the mouth in sheets, which are then punctured and form rings as they are stretched by the airflow. The rings fracture, leaving filaments. Little beads of fluid form on the filaments, which elongate and fragment to finally produce droplets.

Bourouiba was surprised to find so much happening to the fluid outside the mouth — it countered the prevailing assumption that droplets exit the mouth fully formed. To Gerardo Chowell, a mathematical epidemiologist at Georgia State University in Atlanta, this is an important finding because it means that droplet formation could be strongly influenced by environmental conditions such as humidity and temperature. And that could help to explain why some diseases, such as flu, tend to occur more frequently at certain times of the year, he adds, perhaps because the ambient conditions favour the spread and survival of certain microbes.

Bourouiba's research advances previous work measuring sneeze and cough droplet sizes, says Ku. Fluid particles can travel varying distances

➔ NATURE.COM
See Lydia Bourouiba discuss the physics of the sneeze at:
go.nature.com/ysagdr



COURTESY OF L. BOUROUBA/MIT

Lydia Bourouiba of the Massachusetts Institute of Technology in Cambridge uses mathematical modelling to study how sneeze droplets travel.

depending on a lot of different parameters, he says. “If I just tell you the size of the particles, I can’t tell you where they’re going to go. Her work actually shows where they go, with a real sneeze.”

THE NEXT LEVEL

A back injury last year has curtailed some of Bourouiba’s more ambitious outdoor activities. But at work, she and her team are preparing to move into a newly built lab with a biosafety-level-2+ containment room, which will allow them to study the sneezes and coughs not just of healthy participants, but of people infected with colds and flu. In preparation for those studies, she has hired a microbiologist who can help the team to determine the microbial load in droplets and how long pathogens survive in the air or on surfaces while maintaining their ability to infect.

Answering this question will be crucial, says Hughes. “We need to learn more about the concentration of microbes in droplets of varying sizes and the infectious doses of a lot of these pathogens.” The containment room will also allow Bourouiba to control the airflow, temperature and humidity so that she can explore the behaviour of emitted droplets in environments that mimic hospitals, aeroplanes or the tropics.

Bourouiba’s ultimate aim is to compile all of her data into a mathematical model that could be used by public-health officials to identify the most likely routes of transmission and how to reduce the risk of disease spread. The model would suggest, for example, whether the biggest risk of contamination is from the air or from surfaces, or how to change the airflow or temperature to minimize the risk in a hospital. It could predict whether a particular person is at high risk of being a ‘superspreader’ and should be quickly placed in a containment unit. During an emergency situation, when a new disease is spreading but it’s not clear how, it might also help officials to identify the most dangerous environments, such as aeroplanes, so that people can avoid them. Then, as the first infected patients are tested and more is learned about the pathogen, those data could be incorporated into the model to refine the risk assessment.

Chowell, who models the spread of infectious diseases, hopes that Bourouiba’s work could eventually be used to give diseases an ‘airborne score’. Knowing that a pathogen is transmitted by airborne aerosols, say, 85% of the time could give public-health officials a better idea of how fast

and far an outbreak will grow, compared with one that’s just 5% airborne, he says. “Models require data, and I think the efforts of Bourouiba and others will help us better calibrate the design of these models, and this will have an impact on our ability to forecast disease spread in real time.”

That may depend on the disease, however. Work from Donald Milton, an environmental-health scientist at the University of Maryland School of Public Health in College Park, suggests that Bourouiba’s approach may not have much impact on the study of influenza because people with flu rarely sneeze⁵. Studying people with the common cold might be more fruitful, he says, because they sneeze more often.

Milton also cautions that focusing on sneezes and coughs may not capture the whole story of respiratory-disease transmission. Breathing and talking are important to consider as well. He and his team have detected flu viral RNA in particles that were simply exhaled by patients, and they have even cultured viruses from such particles. Bourouiba says that she can study breathing emissions using her methods if they turn out to be a factor, but she first wants to study infected people to see which are the most important emissions to examine.

One occupational hazard for Bourouiba is that it’s difficult to escape her work: whenever she hears a sneeze on a plane or in the classroom, she can’t help thinking about the droplets flying through the air. There is not much she can do about that, but it does remind her why she became so fascinated with fluid mechanics when she was an undergraduate: fluids are everywhere.

Her videos might earn her the nickname of ‘the sneeze lady’, a student once warned her. But she says she doesn’t mind. “If people get interested in the topic because of the humorous aspect, I have no problem with that.” ■

Corie Lok is an editor for *Nature* in Boston, Massachusetts.

1. Yu, I. T. S. *et al.* *N. Engl. J. Med.* **350**, 1731–1739 (2004).
2. Osterholm, M. T. *et al.* *mBio* **6**, e00137-15 (2015).
3. Bourouiba, L., Dehandschoewercker, E. & Bush, J. W. M. *J. Fluid Mech.* **745**, 537–563 (2014).
4. Scharfman, B. E., Techet, A. H., Bush, J. W. M. & Bourouiba, L. *Exp. Fluids* **57**, 24 (2016).
5. Milton, D. K., Fabian, M. P., Cowling, B. J., Grantham, M. L. & McDevitt, J. J. *PLoS Pathog.* **9**, e1003205 (2013).



COMMENT

CONSERVATION UNESCO's marine-heritage arm calls for cash **p.30**

ASTRONOMY A survey of the history, technology and design of telescopes **p.34**

INNOVATION How hippies repurposed science and made it cool **p.36**

OBITUARY Walter Kohn, quantum-chemistry Nobel winner, remembered **p.38**

AMGEN



Researchers prepare to filter crystals at an Amgen small-molecule-manufacturing facility.

Drug companies must adopt green chemistry

John L. Tucker and Margaret M. Faul describe how they transformed their company to save time and money by making drugs sustainably.

In the past decade, many large pharmaceutical companies have moved to using green-chemistry practices for drug discovery, development and manufacturing. These firms include ours, Amgen, and others such as the Merck Group, Abbott, Johnson & Johnson and Roche. Ranking systems such as the Dow Jones Sustainability Indices and the Pacific Sustainability Index¹ track how well firms are doing. This shift is being driven by the realization that processes that

are cheaper and environmentally superior deliver a competitive advantage.

Success depends on instilling a culture of sustainability into a firm. Staff at all levels — from management to lab scientists — need to understand the concepts of green chemistry and how they might be embraced to everyone's benefit. The future rewards of major operational changes need to be visualized and funding must be put in place before any results are realized.

Here we set out how to build such a culture, based on our experience at Amgen, which is headquartered in Thousand Oaks, California. Change cannot be achieved overnight, but we found that a series of small wins can build momentum until the vision becomes clear and is accepted. By following a combination of bottom-up² and top-down³ approaches, Amgen rose from having a C+ rating on the Pacific Sustainability Index in 2007 to become one of the top-rated (A+) companies in ►

► 2012 (see 'Sustainability scores'). The ingredients are: an empowered team with management support to lead the transformation; staff education, awareness and recognition; investment in technology; development of metrics and tools; and external collaboration and outreach^{4,5}.

Our approach focuses on the 'triple bottom line'⁶ — profit, people and planet — and the 12 principles of green chemistry⁷, which include minimizing ingredients, waste, toxicity and energy. Each of these principles can be applied to a sector or product. For instance, a detergent's formula might be redesigned so that it degrades without accumulating, persisting or releasing toxic chemicals into the environment.

Drug development has its own challenges. A pharmaceutical's therapeutic response depends on its molecular structure. Existing drugs have already survived a battery of toxicological and therapeutic tweaks, as well as clinical studies that took years to complete. Redesigning a drug to be more readily degradable could become an entirely new development project — changing its structure might alter its function. But many green-chemistry principles can make processes more efficient, for example by reducing the number of reaction steps, or by using less energy or materials (see 'Green-chemistry principles to drive sustainability').

The challenge is to catalyse a new norm across the industry while meeting institutional expectations and demonstrating value.

SEVEN STEPS

Empower champions. The first step taken at Amgen was to create a high-level green-chemistry team that spanned the company's many functions. The aim was to define and entrench green-chemistry expectations across the company — framed as 'how, why, what and where'. The team initially comprised six scientists — a chair (one of us, J.L.T.), representatives from process and analytical chemistry, engineering, environment, health and safety, and drug-production technologies — and was supported by an executive sponsor (the other of us, M.M.F.). Medicinal-chemistry teams and biological-molecule representatives were added later. All steps of drug discovery, development and manufacturing at Amgen were considered across the company's multiple sites. Resources had to be secured for communication, collaboration and development of green methodologies and technologies.

Raise awareness. The green-chemistry team at Amgen set up a series of lectures by innovators and thought-leaders from within and outside the company to spread knowledge of green-chemistry principles, their potential for drug development and

examples of good practice. Websites and visuals were created and disseminated. For example, notices highlighting the 12 principles were placed on hoods and in laboratories, and green-solvent and reagent-selection guides were supplied. Challenges had to be acknowledged, including practical difficulties in using sustainable methods and materials, the importance of maintaining industry competitiveness during the transition, as well as the need for transparency in meeting regulatory expectations.

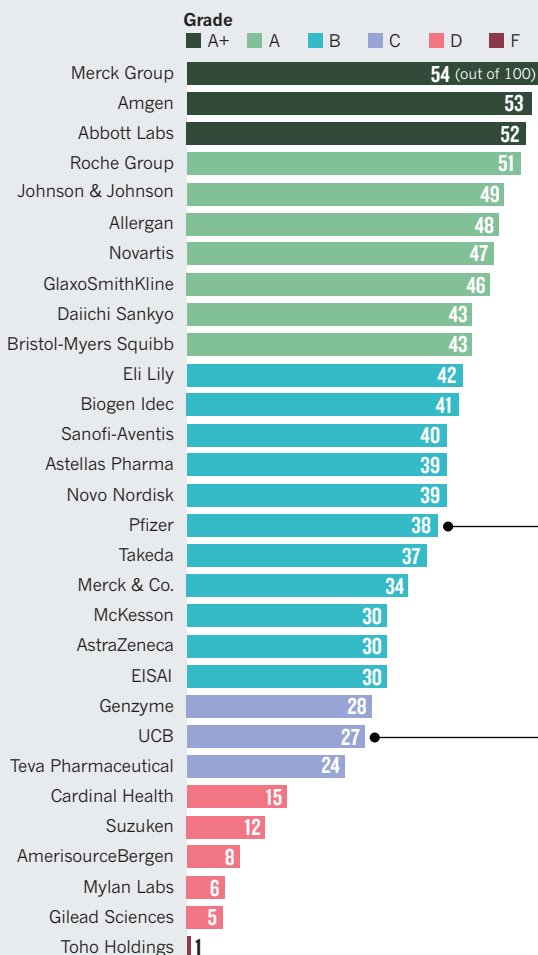
The greatest difficulty we encountered was providing a vision of the future state of the organization and convincing people that it was achievable and worth pursuing. Employees needed strong incentives to adopt sustainable principles in their already challenging careers. Productivity and operational efficiency were expected to rise regardless of change, and developing new drugs across the industry was becoming increasingly difficult. Converting sceptics was hard. Scientists demanded data,

precedents and straight answers as to why or how an initiative should or should not be pursued. Business leaders wanted figures on efficiency, cost and environmental impact.

Expand collaborations. To share knowledge, Amgen joined the American Chemical Society's Green Chemistry Institute's Pharmaceutical Roundtable and the IQ Consortium's Green Chemistry Working Group. These interactions provided: industry harmonized tools, such as solvent and reagent guides; insight into industrial green-chemistry strategies and practices; access to metrics; opportunities to influence academic research; and discussions with governmental agencies such as the US Environmental Protection Agency, the Food and Drug Administration⁸ and the National Science Foundation. Getting many companies and regulators around the same table helped to smooth the adoption of green-chemistry principles, ensuring that cheap, safe and fast access to new therapies would continue.

SUSTAINABILITY SCORES

The Pacific Sustainability Index rates information reported on the websites of the 20 largest drug companies. Scores out of 100 are given in areas including environmental and social measures, and are based on a standard questionnaire. The top 4% receive A+ ratings; the bottom 4%, F.



Amgen scored highly in social and environmental intent by reporting on areas that it still needs to improve, such as water and air emissions.

Pfizer posted a lot of information, but more specifics on topics such as employee relations would have raised its scores.

UCB published its first Corporate Social Responsibility Report in 2011 that stated goals for improvement.

2012 index based on companies in the Forbes 2010 Drugs and Biotechnology sector list.

SOURCE: REF. 1

Define metrics. We collated and shared quantitative tools for identifying internal strengths and weaknesses and best practices. Large pharmaceutical firms track the efficiency and waste of their drug portfolios through metrics such as the E factor, which measures kilograms of waste generated per kilogram of product, or process mass intensity (PMI), a similar metric that measures the total mass of materials per mass of product. The Amgen green-chemistry team developed metric calculators for the electronic notebooks used by scientists in the lab to pinpoint the most inefficient steps and operations within projects and to compare progress as a project moves through process development.

The metrics exposed a clear correlation between reducing waste and lowering process costs across Amgen. For example, using fewer materials in one synthetic process that was being developed for a new clinical candidate reduced the E factor by 82% and the cost by 83% while also improving the overall yield.

Recognize achievements. An internal award programme raised awareness among employees of the impact and advantages that green chemistry brings. The competition sought and broadcast the best examples of how adopting the principles led to better outcomes. Leaders presented prizes and the winners served to inspire others. For example, researchers working on etelcalcetide (which is being assessed as a treatment for a complication of chronic kidney disease) received Amgen's green-chemistry award for developing a process that reduced their organic-solvent use by more than 400,000 litres and shortened the projected manufacturing processing time.

Invest in technology. We explored new methods. For example, we switched to using enzymes as reagents in the synthesis of small molecules, moving away from convention reactions catalysed by transition metals. This reduced the number of steps and increased reaction throughput. For example, enzyme catalysis allowed us to make a key fragment of a drug candidate in early development, reducing the time to manufacture by 80%. It also eliminated⁴ volumes of organic solvent used during chromatographic purification of small molecules, doubled the yield and reduced the cost of the starting material by more than 99%.

A standard procedure for oxidizing double bonds to aldehydes is ozonolysis (aldehydes are key chemical groups that enable the assembly of molecules). In one case study, our initial synthesis of an aldehyde intermediate used a flammable solvent in the presence of oxygen. Using a 'continuous flow' process — in

GREEN-CHEMISTRY PRINCIPLES TO DRIVE SUSTAINABILITY

CONCEPT	GOOD FOR PLANET	GOOD FOR PROFIT
Atom economy	Fewer by-products.	More value from less material.
Minimize solvent use	Less waste, less energy.	Higher throughput.
Optimize reagents	Recyclable reagents and catalysts minimize volume of chemicals needed.	Higher efficiency.
Convergent synthesis (produce several pieces of a molecule at once)	Increases efficiency, saves energy.	Higher efficiency, fewer operations.
Reduce energy use	Less pollution from power generation and transport.	Shorter, more efficient processes under milder conditions.
Analyse reactions in real time	Reduces exposure or release to environment.	Increases throughput and process efficiency, fewer reworks.
Prioritize safety	Non-hazardous materials reduce risk of exposure, release, explosions and fires.	Reduces potential harm to workers, down time and need for special control measures.

which the conditions of a stirred reactor are optimized in real time to maximize the yield — rather than batch mode allowed us to process large quantities of material in a short time (5 kilograms of aldehyde intermediate in 18 hours) without building up dangerous amounts of reagents, intermediates, solvents, ozone and oxygen gas.

Even simple changes, such as using large, disposable plastic rather than stainless steel vessels for manufacturing biologic drugs (made using recombinant DNA technology), saved time, space, effort and money. Although it creates more solid waste, single-use vessels do not need rooms and resources such as water or steam to clean or sterilize them. Costs fall and production capacity can grow without increasing — and even by decreasing — the waste footprint of the plant.

Promote outreach. To spread the sustainability mindset across the industry, it will be crucial to work with academics to prepare the next generation of green chemists and with regulators to assess and reward efforts. To this end, Amgen scientists regularly give talks in universities on green chemistry. The public and investors value sustainable practices (albeit in ways that are hard to quantify) so it is important that they know of the company's commitment. To this end, corporate reporting must be transparent and present an unvarnished and accurate view.

All this hard work has paid off. In 2013, Amgen was declared a top performer in the pharmaceutical industrial segment by the Dow Jones Sustainability Index, and placed 21st in *Newsweek's* 2015 ranking of green US companies from all sectors.

The financial benefits are already clear.

Sustainable Asset Management, a company that directs investment dollars, has named Amgen a 'sustainability mover', opening up new sources of investment.

Amgen's initiative is not window dressing. It is rooted deep in a broad commitment to deliver medicines to people in a way that uses fewer resources and promotes industry competitiveness. We will continue to focus on the principles of green chemistry, seek operational efficiency, explore new technologies and support education and research. And we will work with other companies in non-competitive areas to encourage the spread of green chemistry throughout the industry. There is still much to do to convince others to create a culture of sustainability.

Green chemistry can deliver for people, planet and profit. Those who embrace it will reap the benefits in future. Those who fail to evolve may cease to be relevant. ■

John L. Tucker is a senior scientist in process development, and **Margaret M. Faul** is executive director of process development, at Amgen Inc., Thousand Oaks, California, USA.
e-mail: tuckerj@amgen.com;
mfaul@amgen.com

1. Morhardt, J. E. *et al.* 2012 Sustainability Reporting of the World's Largest Drugs and Biotech Companies (*Pharmaceuticals) (Roberts Environmental Center, 2012); available at go.nature.com/su9ogu
2. Fraser, E. D. G., Dougill, A. J., Mabee, W. E., Reed, M. & McAlpine, P. J. *Environ. Mgmt* **78**, 114–127 (2006).
3. Tucker, J. L. *Org. Process Res. Dev.* **14**, 328–331 (2010).
4. Leahy, D. K. *et al.* *Org. Process Res. Dev.* **17**, 1099–1109 (2013).
5. Tucker, J. L. *Aldrichimica Acta* **48**, 16–17 (2015).
6. Elkington, J. *Cannibals with Forks: The Triple Bottom Line of 21st Century Business* (New Society, 1998).
7. Anastas, P. T. & Warner, J. C. *Green Chemistry: Theory and Practice* (Oxford Univ. Press, 1998).
8. Ritter, S. K. *Chem. Eng. News* **92**, 32–33 (2014).



The seas cannot be saved on a budget of breadcrumbs

The marine arm of UNESCO's World Heritage Convention needs secure funding to realize its vast potential to protect the ocean, argues **Fanny Douvere**.

A global exemplar in protecting Earth's most iconic places is the 1972 World Heritage Convention of the United Nations Educational, Scientific and Cultural Organization (UNESCO)¹. The convention's founding years are credited with rescuing the ancient Egyptian temples of Abu Simbel from being lost under the Nile — an operation that required collaboration across more than 50 countries.

World-heritage recognition has since become a hallmark for sustainable protection of valuable sites, from Peru's Machu Picchu to Tanzania's Serengeti National Park. UNESCO's World Heritage List reflects the common heritage of humankind, a legacy to pass on to future generations. But its impact is felt mostly on land.

UNESCO also has a World Heritage Marine Programme, which I lead. It was created by UNESCO's World Heritage

Committee just over ten years ago to help secure effective conservation for marine sites on this list. Although the programme has a powerful brand that enables effective negotiation with government bodies and civil society, it is unfunded. Like a non-governmental organization (NGO), it must secure financial support from various sources. Finding these sources is difficult. Governments struggle to cover the costs of specialized programmes across the UN. Philanthropic organizations often seem more comfortable funding research institutions or NGOs. The World Heritage Marine Programme currently has just 3 professionals to cover 47 sites across 36 nations (see 'Ocean treasures').

World-heritage marine work can help governments to design feasible approaches to the threats that face some of the last wild places on Earth. Despite UNESCO's established ability to influence governments, and

to formulate and implement effective change for sustainable management, philanthropic contributions are hard to come by. With the oceans facing existential threats from pollution, climate change and overfishing, it is time to invest in one of the best tools available for conservation.

TRACK RECORD

Our programme can deliver far-reaching impacts where others cannot.

In 2011, the Australian government considered the protection of the Great Barrier Reef adequate even as scientists increasingly warned that the reef was in poor condition and getting worse^{2,3}. Despite the iconic status of the reef, and it having been a prime example of marine-protected-area management for 40 years, the site had suffered from decades of incremental decisions that threatened 'death by a thousand cuts'. More than two-thirds of



Australia's Great Barrier Reef is a UNESCO World Heritage Marine Site.

XL CATLIN SEAVIEW SURVEY

all coastal-development proposals near the reef submitted between 1999 and 2011 had been approved. Previous government financial commitments to halt and reverse the declines in water quality — declines largely responsible for the loss of coral systems closest to the coast — came up for revision in 2013 but renewal was uncertain. In 2012, the World Heritage Committee issued its first warning that it would list the site as 'world heritage in danger' unless it saw proof of substantial progress by the following year.

This opened the way for the UNESCO World Heritage Centre and its scientific advisory body, the International Union for Conservation of Nature, to embark on extensive negotiations with the Australian government, eventually changing the government's approach entirely. It reversed its original plan to dump 3 million cubic metres of dredged material into the Great Barrier Reef. Indeed, in 2015, it banned the dumping of dredged material throughout the world-heritage site — an area larger than Italy. Australia's government committed more than Aus\$200 million (US\$145 million) to improve water quality and set an ambitious aim to reduce pollution runoff by 80% by 2025. Proposed port-development areas have been restricted from 11 to 4 major ones, and future coastal development must align with a strategic plan aimed at improving the

health of the reef between now and 2050.

In his address to the World Heritage Committee last July, Australia's environment minister, Greg Hunt, said that UNESCO advice had allowed Australia "to do in 18 months what otherwise would have taken decades".

Something similar happened in the Belize Barrier Reef, the world's second largest coral-reef system. This site was placed on the list of world heritage in danger in 2009 because of the destruction of mangrove forests for coastal development and ongoing threats of offshore oil exploration. We began intensive talks with the government and stakeholders in early 2015, which led to a road map to reverse the endangered status. Last December, following years of deadlock, the Belizean government announced a permanent ban on all oil exploration in the site; in February, it approved an ambitious coastal-management plan. These are concrete steps that can lead to a brighter future for this unique array of reef types, and the nearly 200,000 Belizeans who depend on it for their livelihoods.

These are just tasters of the sustainable change that a properly funded World Heritage Marine Programme could bring. Since the first true marine site was inscribed on the UNESCO World Heritage List in 1982, our scope has grown into a global collection of sites that stretch from the tropics to the Arctic. The list includes the breeding grounds of the world's last healthy population of grey whales (*Eschrichtius robustus*), in Mexico; the highest density of ancestral polar-bear dens, in Russia; and the home of one of the world's most ancient fish, the coelacanth, in South Africa, and that of the inimitable marine iguanas (*Amblyrhynchus cristatus*), in the Galapagos Islands.

EFFECTIVE MANAGEMENT

Most of these places host a range of activities aimed at conservation and at generating income. Tension between opposing concerns is inevitable, and the most durable solutions emerge when diverse viewpoints of activists, scientists and government officials are effectively mediated. Our programme is uniquely positioned to take on this role.

In its first few years, the World Heritage Marine Programme had limited capacity and mainly worked by supplying recommendations and basic guidance to a handful of sites. But better science, the uncertainties of climate change and increasing pressure to use ocean space have brought shifting information and demands. Effective management of the flagship marine protected areas required us to adopt a more flexible, hands-on approach. Now we coordinate technical support missions, bring site managers and external experts together to exchange ideas, and increasingly broker solutions with government leaders to secure the urgently needed protection of

irreplaceable marine ecosystems.

Our work in Belize and Australia shows that if we can dig into the nitty-gritty of problems collaboratively we can affect change. But without a regular income, the programme cannot support even the sites that most need focused attention. For every case like the Great Barrier Reef, there are other urgent ones that do not get support.

For example, for nearly a decade, the World Heritage Committee has recommended that the Panamanian government establish a comprehensive plan for sustainable fisheries in Coiba National Park. Overfishing in

this park, considered the jewel of Panama, has almost wiped out once-abundant hammerhead sharks and boosted the jellyfish population. We must take advantage of the possibility of

working with the Panamanian government to deliver sustainable fisheries and secure a healthy future for Coiba.

The same can be said for the Sundarbans in Bangladesh, part of the largest mangrove forest in the world, which is now threatened by coastal development. Another example is Banc d'Arguin National Park in Mauritania, where catch from fishing just outside the site's borders has increased more than 12-fold from 1994 to 2010. But with current funds, we simply cannot be everywhere.

OCEAN INVESTMENT

To address the challenges at Coiba and other sites that struggle with illegal or unsustainable pressures, the World Heritage Marine Programme needs to be recognized as an effective body worthy of investment. It must be able to plan for the long term and focus attention on the most urgent priorities.

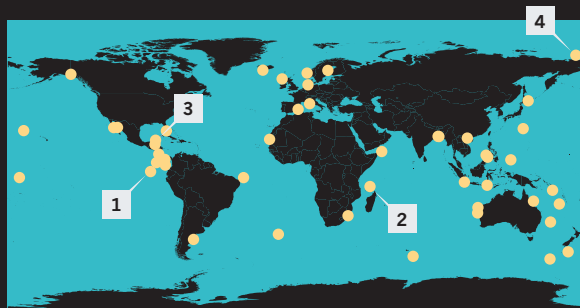
The successes mentioned here were made possible largely as a result of the steady and mostly unconditional support that Swiss watchmaker Jaeger-LeCoultre has provided to the programme since 2008. This has allowed the programme to steer away from a short-term project-by-project approach, and instead concentrate on the type of input needed to achieve the ultimate measure of success: improved conservation of sites' treasured values that won their world-heritage recognition in the first place.

To expand efforts and coordinate our work so that it is efficient and impactful, we need a broad base of stable financial support.

We now stand at a moment of even greater opportunity to preserve the open ocean — waters not subject to any single country's jurisdiction. These expanses, known as the high seas, cover half our planet. They also need protection that few — if any — mechanisms provide. From 2010 to 2012, seven ►

Ocean treasures

UNESCO recognizes 47 exceptional world-heritage marine sites across 36 nations — from mangrove forests in Belize to polar-bear breeding grounds in the Arctic.



1 Marine iguana (*Amblyrhynchus cristatus*) in the Galapagos Islands.



3 Mangroves in the Belize Barrier Reef.



2 Green turtle (*Chelonia mydas*) in Aldabra Atoll.



4 Polar bears (*Ursus maritimus*) in the Wrangel Island Reserve system.



► marine protected areas were established, covering more than 285,000 square kilometres in the Atlantic Ocean under the auspices of OSPAR, a cooperative effort by 15 governments and the European Union to protect the northeast Atlantic. These are some of the first of very few protected areas in the high seas. But this is only a regional action. The UN, under the 1982 Convention on the Law of the Sea, has started negotiations for a possible new agreement to protect high-seas biodiversity on a global scale. This framework agreement is real progress, but as yet lacks practical procedures to nominate, oversee and protect sites.

Enforcing protection of the high seas is one of today's biggest challenges in ocean conservation. The world-heritage system is equipped to help get this protection in place. It has a 40-year history of identifying and overseeing the state of conservation of places of Outstanding Universal Value across 163 nations and has had ample successes. Such institutional experience is unparalleled in nature conservation, but its capacity to navigate such complexities and to preserve ecosystems is often overlooked.

In a February interview with *The New York Times*, biologist E. O. Wilson called for creating something equivalent to the UN world-heritage sites to protect the open ocean as priceless asset of humanity.

Our ability to engage constructively with government is starting to produce real, lasting results in conservation. It could be replicated in other marine sites to great effect. Strengthening the international oversight of flagship protected areas will amplify the work of scientists, local NGOs and related organizations. Their concerns become international causes that, through a tactical and skilled use of the World Heritage Convention, can lead to government action that benefits us all.

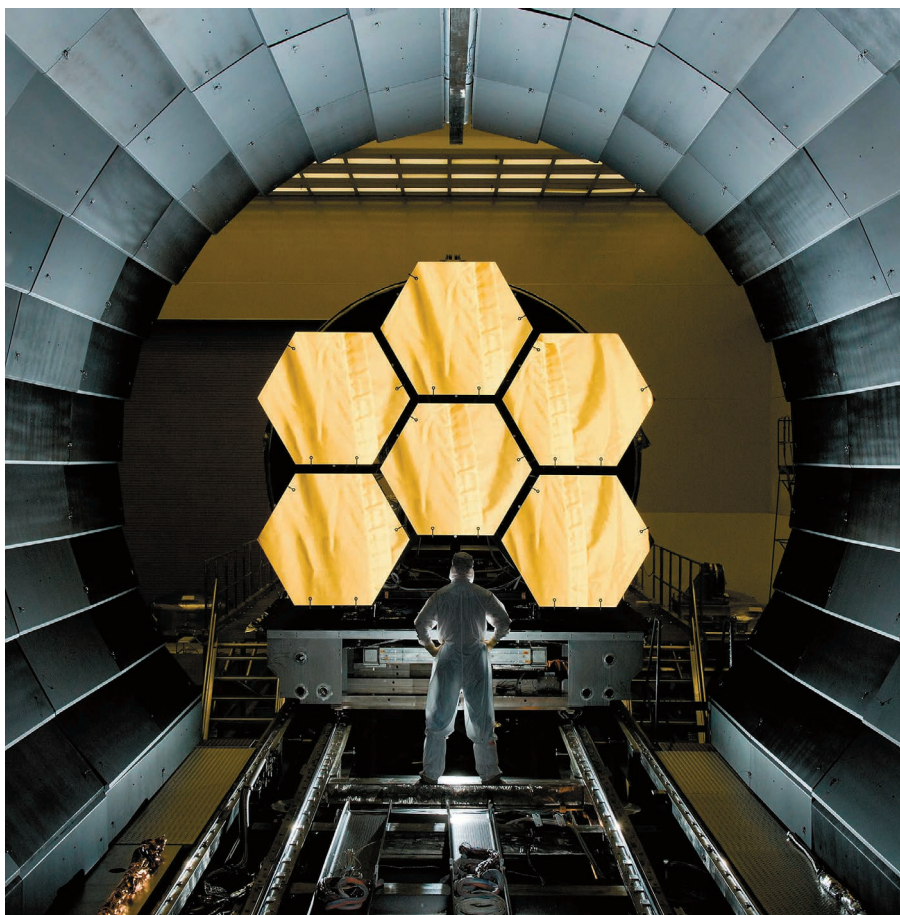
Being effective requires robust investment. The World Heritage Convention cannot change the world on a budget of breadcrumbs. Philanthropists seeking investments that make lasting changes should look beyond their conventional outlets of NGOs and research institutions and consider this potential. Ignoring world heritage is a lost chance for our oceans. ■

Fanny Douvère is coordinator of the World Heritage Marine Programme of the United Nations Educational, Scientific and Cultural Organization (UNESCO), Paris, France.
e-mail: f.douvere@unesco.org

1. Cameron, C. & Rössler, M. *Many Voices, One Vision: The Early Years of the World Heritage Convention* (Routledge, 2013).
2. Hoegh-Guldberg, O. *Mar. Freshw. Res.* **50**, 839–866 (1999).
3. Hoegh-Guldberg, O. et al. *Science* **318**, 1737–1742 (2007).

The views expressed are those of the author and do not necessarily represent those of UNESCO.

IGUANA: TUI DE ROY/MINDEN PICTURES/FLPA; TURTLE: WIL MEINDERTS/MINDEN PICTURES/FLPA; MANGROVES: BRIAN J. SKERRY/NATIONAL GEOGRAPHIC/GETTY; POLAR BEARS: SERGEY GORSHKOV/MINDEN PICTURES/GETTY



The primary mirror of the James Webb Space Telescope, which will launch in 2018.

ASTRONOMY

Cosmic detectives

Bernie Fanaroff surveys a study that probes telescopes in history and across the electromagnetic spectrum.

It is a little odd that many astronomers still call themselves optical, radio or X-ray astronomers. The major problems of astrophysics and cosmology, such as how stars form and the nature of active galactic nuclei, cannot be solved by observing in only one part of the electromagnetic spectrum. Thus we live in the era of multi-wavelength and multimessenger astronomy, which demand different kinds of telescope and technology to observe different parts of the spectrum and even other particles and waves, such as neutrinos, cosmic rays and gravitational waves.

In *Eyes on the Sky*, British astronomer Francis Graham-Smith delivers a valuable survey of the history, technology and design of telescopes across the electromagnetic spectrum, starting with Galileo Galilei's pioneering seventeenth-century refracting

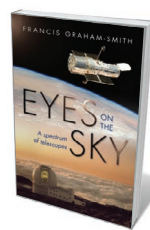
telescope. Graham-Smith explains the principles of how telescopes, such as optical reflectors or X-ray telescopes in space, make images or spectra and how they detect waves and photons, using everything from radio receivers to solid-state mega-pixel charge-coupled devices. As he notes, the field has been transformed, especially in recent years, through a combination of technical advances and radical change in astronomy's organization and scale, with the advent of large international teams and multinational projects.

Astronomers are always pushing the boundaries of technology, out of the need to detect more and more of the spectrum from increasingly faint objects. Graham-Smith's account of that process is fascinating.

NATURE.COM

For more on science in culture see: nature.com/booksandarts

Some of the groundbreaking technological advances have been in detectors, very fast electronics and computing, and space telescopes. As he explains, satellite observation is essential for the parts of the spectrum blocked by Earth's atmosphere, such as X-rays. New-generation telescopes of this kind currently include NASA's Kepler



Eyes on the Sky: A Spectrum of Telescopes
FRANCIS GRAHAM-SMITH
Oxford University Press: 2016.

space observatory and the European Space Agency's Planck satellite telescope, while next-generation satellite instruments will include NASA's James Webb Space Telescope (JWST). And the JWST, along with ground-based instruments such as the radio telescopes of the Square Kilometre Array (SKA) in South Africa and Australia, will produce huge quantities of data from sky surveys of unimagined sensitivity and scope. With young researchers able to access a flood of wonderfully exciting data, this will be a new golden age. Meanwhile, the detection of gravitational waves with the Advanced Laser Interferometer Gravitational-Wave Observatory, announced this year (S. Rowan *Nature* 532, 28–29; 2016), heralds the beginning of gravitational-wave astronomy.

Graham-Smith gives a useful summary of what is to come from these telescopes and surveys. Huge data sets of galaxies and other objects are being produced by sky surveys at different wavelengths, and many astronomers spend a large part of their time cross-matching the objects found. For instance, objects found in radio surveys must be matched with those found in surveys at optical wavelengths, to learn about the source of the radiation (such as a galaxy) and its distance (measured using the Doppler shift, or 'red shift', of the spectrum, which is caused by the expansion of the Universe).

There are many challenging problems ripe for cracking. One is the structure of the Universe. How did a once-uniform ball of very hot gas and energy become a highly structured, complex Universe, evolving over the 13.8 billion years since the Big Bang? Optical surveys such as the Sloan Digital Sky Survey tell us about the distribution of galaxies, galactic clusters and super-clusters. Minute fluctuations in the cosmic microwave background radiation — measured by the Planck telescope, among others — tell us about conditions when the Universe was only 380,000 years old, before the first stars, galaxies and clusters formed. The SKA will probe this 'cosmological dawn' and track the development of structure by looking at the history of hydrogen gas in the Universe.

As Graham-Smith discusses, the structure and evolution of galaxies is another hot topic. Almost all galaxies have supermassive black holes spinning at their centres, with masses millions to billions of times that of the Sun. Vast amounts of energy are radiated from near the black hole, or carried off as kinetic energy by collimated (very narrow) jets squeezed out along the poles of rotation and extending, in some cases, for a megaparsec. This process is probably key to the formation of stars and the evolution of galaxies. Energy from the jets and radiation is dumped into the gas between the stars and galaxies, and is believed to significantly influence the rate of star formation and, as a result, galaxy evolution. The heating and stirring of the gas in turn affects the rate of accretion and energy generation around the black hole, in a powerful feedback mechanism.

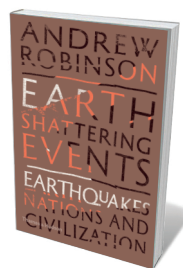
To refine their picture of this activity, astronomers are marshalling findings from a range of telescopes to map the jets' radio emission and estimate their kinetic and magnetic energy, as well as the energy emitted at optical and ultraviolet wavelengths. They are using X-ray observations to determine how hot the gas is, and infrared observations to gauge how much dust there is in the interstellar medium. They observe spectral lines at millimetre wavelengths to map the outflow of molecular gas. X-rays and γ -rays also tell us about the gas dynamics close to the black hole or in the region where the jets are launched.

Eyes on the Sky does contain a few surprising errors. For example, the Karl G. Jansky Very Large Array radio telescope in New Mexico, for instance, still has 27 dishes after its upgrade, not 36. Nevertheless, Graham-Smith's book is a very interesting explanation of the multitude of telescopes and their history.

Telescope technology continues to develop at breakneck speed. The SKA, for instance, demands new technologies to increase sensitivity, process huge quantities of data very fast and keep costs in check. This and other planned great observatories — the JWST, as well as the γ -ray seeking Cherenkov Telescope Array and the optical/near-infrared European Extremely Large Telescope on the ground — are likely to produce major discoveries in areas such as transient sources of radiation, the understanding of planet formation, the nature of dark matter and the history of the Universe. They will undoubtedly also uncover unknown unknowns, those serendipitous discoveries that are the hallmark of the great telescopes of history. ■

Bernie Fanaroff was the director of the Square Kilometre Array South Africa project until the end of 2015 and is now a part-time strategic adviser to the project.
e-mail: bfanaroff@ska.ac.za

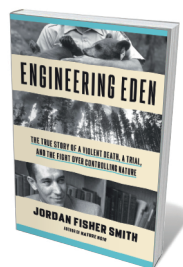
Books in brief



Earth-Shattering Events: Earthquakes, Nations and Civilization

Andrew Robinson THAMES AND HUDSON (2016)

A “fatal attraction”: geophysicist James Jackson’s description of humanity’s penchant for living in earthquake zones is all too apt, notes science writer Andrew Robinson in this compelling history of seismicity and society. Robinson traces more than 2 millennia of cataclysms, vividly evoking events such as the magnitude-8.8 quake-cum-tsunami that largely flattened Lisbon in 1755. Woven through is a history of seismology from its first glimmerings in ancient China, through geologist John Milne’s groundbreaking work in the nineteenth century to today’s hurdle-ridden drive to predict seismic risk.



Engineering Eden

Jordan Fisher Smith CROWN (2016)

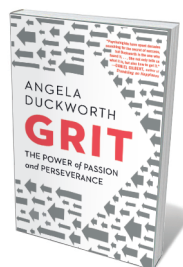
In 1972, a grizzly bear eviscerated tourist Harry Walker in Yellowstone National Park, Wyoming. His family’s lawsuit against the US National Park Service ignited a vastly broader debate about ‘managed nature’. In this beautifully synthesized study, writer (and former ranger) Jordan Fisher Smith argues for symbiotic balance in our interaction with the wild, because “the ties that bind, bind in all directions”. As he shows, expert witnesses such as zoologist Starker Leopold helped to shift Yellowstone’s mismanagement of bears — notably the deliberate feeding that predisposed them to attack.



Blue Skies over Beijing: Economic Growth and the Environment in China

Matthew E. Kahn and Siqi Zheng PRINCETON UNIVERSITY PRESS (2016)

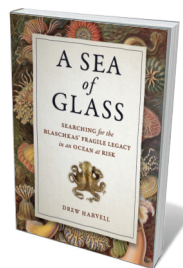
Beijing’s atmospheric pollution in 2013 reached 40 times the safe level set by the World Health Organization. To gauge progress on the country’s urban sustainability, economists Matthew Kahn and Siqi Zheng apply microeconomics to industry, pollution dynamics, and local and central government efficacy. They see that analysis — along with factors such as growing environmental awareness in China, and evidence of sharply improved air quality in some post-industrial US cities — as potentially heralding a turnaround.



Grit: The Power of Passion and Perseverance

Angela Duckworth SCRIBNER (2016)

When psychologist Angela Duckworth received a MacArthur Fellowship, or ‘genius grant’, in 2013, the irony was not lost on her; for years, her father had said she was “no genius”. But Duckworth saw sheer dogged effort as brilliance of a different sort, and ultimately more important to achievement than talent. She lucidly anatomizes the nature of grit, drawing on her own and others’ research (such as psychiatrist George Vaillant’s ‘treadmill test’), and explicating the passion, purpose, practice and optimism that feed perseverance and resilience. A deft corrective to IQ culture.



A Sea of Glass: Searching for the Blaschkas' Fragile Legacy in an Ocean at Risk

Drew Harvell UNIVERSITY OF CALIFORNIA PRESS (2016)

In nineteenth-century Bohemia (now the Czech Republic), master glassblowers Leopold and Rudolph Blaschka spun supremely lifelike replicas of organisms as teaching tools. Ecologist Drew Harvell, finding more than 500 models of marine invertebrates at Cornell University in Ithaca, New York, set out to restore them. Stunning photos of a number of them contextualize the dramatic taxonomic and ecological shifts in ocean life over the past 150 years. [Barbara Kiser](#)

HISTORY

Peace, love and lab work

Ann Finkbeiner delves into a collection reappraising the hippy tech-heads, agronomic groovers and far-out ecodesigners of the 'long 1970s'.

For a lively decade or so in the 1960s and 1970s, the younger generation in the United States looked at its elders — with their unwinnable wars, florid military-industrial complex, intransigent racism and contaminated brownfields — and was outraged. In particular, young people rejected what they saw as the foundations of many establishment ills: the weapons and toxic chemicals spawned by science and technology. That is the standard history, say historians of science David Kaiser and Patrick McCray — and it's not quite right.

In their edited volume *Groovy Science*, Kaiser (author of *How the Hippies Saved Physics* (W. W. Norton, 2011); see H. Gusterson *Nature* **476**, 278–279; 2011) and McCray show that in the “long 1970s”, the young, in creating a counterculture, didn't so much reject science as recreate it. Each essay is a case history on how the hippies repurposed science and made it cool.

What they rejected was the work of defence contractors, big government or corporate labs, which they deemed hierarchical, inflexible and bound to special interests. By contrast, ‘groovy’ science was (as hinted at by the word's origins in 1930s jazz) playful and improvisational, small-scale and done in the name of peace by “world-thinkers, dropouts from specialization”. Their research ranged from the practical (light, strong surfboards) to the visionary (space travel). Because some were drug-addled, it also encompassed the hare-brained (communication with dolphins, the fervent wish of physician John Lilly). Some of it is now thoroughly mainstream.

A number of these hippies were conventionally trained scientists with doctorates. Psychologist Abraham Maslow looked beyond the behaviourism — the idea that humans are blank slates who react to stimuli — advocated by psychologist B. F. Skinner and others. Maslow's alternative was a humanistic ‘hierarchy of needs’, the fulfilment of which would lead to happiness, self-actualization, and ultimately a better society. He became a patron of the Esalen Institute in Big Sur, California, a centre for workshops, encounter groups — forums that encouraged face-to-face communication and confrontation — and yoga classes, all aimed at training people to realize their potentials. His innovative approach, focusing on good mental health rather than pathological symptoms,



Writer Stewart Brand in 1966.

persists in ‘positive psychology’.

John Todd, a biologist at San Diego State University in California, worried that industrial agriculture was creating dangerously limited monoculture. He organized a network of like-minded professionals, the New Alchemists, who encouraged agriculturally self-sufficient communities and built enclosed ecosystems, or ‘bioshelters’, such as the Ark on Canada's Prince Edward Island. Similar principles and technologies are now used by ecodesigners in creating green buildings that incorporate renewable materials and have sustainable energy demands.

Countercultural researchers jump-started interest in midwifery and home births; they also learned to produce their own cheese, making goat's cheese “no longer weird” in the United States. The grooviest of them all were arguably the bricoleurs, engineers in garages, who used “whatever comes to hand”. As groundbreaking publisher Stewart Brand wrote in the *Whole Earth Catalog* (part encyclopaedia and part



Groovy Science: Knowledge, Innovation, and American Counterculture
EDITED BY DAVID KAISER AND W. PATRICK MCCRAY
University of Chicago Press: 2016.

how-to manual, often cited as anticipating the Internet), these were doers “with a functional grimy grasp on the world”. Welder and designer Steve Baer and his friends experimented with passive solar collectors, and geodesic domes partly crafted from junked cars, in Colorado commune Drop City. He adapted the ideas for his New Mexico-based company Zomeworks. James Baldwin, a jack-of-all-trades, built a truck with fold-out sides that served as a travelling workshop and classroom for ecodesign technologies such as solar panels. Baldwin's toolkit was used to construct sustainable buildings, including the Farallones Institute in Occidental, California, established by architect Sim Van der Ryn to teach ecological design. As with the New Alchemists, this stream flowed into contemporary ecodesign.

Most noticeable about these science freaks was their cheery willingness to share whatever they knew or learned through manuals, periodicals and books, many of them best-sellers. Publications such as the magazine *The Great Speckled Bird* spread the ecodesign gospel. Science and science-fiction magazine *Omni*, influenced by psychologist and avid user of hallucinogenic drugs Timothy Leary, popularized his ideas of space migration and transhumanism — transcending human limitations. The *Whole Earth Catalog* famously offered tools for developing “person power” — everything from hammers to guides for building a pipe organ and books on population control. “We are as gods,” Brand wrote, “and might as well get good at it.”

For the academic historian, *Groovy Science* establishes the “deep mark on American culture” made by the countercultural innovators. For the non-historian, the book reads as if it were infected by the hippies' democratic intent: no jargon, few convoluted sentences, clear arguments and a sense of delight. Because of “acid-spooked scientists, stoned tinkerers, and many of the other straight-up hippies and freaks,” write historians Beth Bailey and David Farber in the afterword, “a substantial subset of Americans came to rethink how they eat, how they communicate, how they stay healthy, how they design and build, and how they have fun.” ■

Ann Finkbeiner is a science writer in Baltimore, Maryland.
e-mail: anniekf@gmail.com

ZEITGEIST/EVERETT/REX/SHUTTERSTOCK

Correspondence

Broad Institute keeps CRISPR tools open

As chief communications officer at the Broad Institute of MIT and Harvard, I wish to clarify that the institute makes patent rights for CRISPR–Cas9 genome-editing technologies available globally across academia and industry (see J. Sherkow *Nature* **532**, 172–173; 2016).

For academic research, the patent rights are freely available and we openly share CRISPR reagents through the non-profit repository Addgene. So far, Addgene has processed more than 30,000 requests for these reagents.

For commercial research, we designed a non-exclusive licensing model. For commercial products, we also follow a non-exclusive model — except for human therapeutics, for which we use an ‘inclusive innovation model’. This is because companies often need exclusivity to justify investing in expensive clinical trials.

The CRISPR–Cas9 licensing agreement with our primary licensee, Editas, stipulates that, for target genes not being pursued by Editas, we (Broad, Harvard and MIT) will make the licences available to other parties to develop new medicines. This helps to ensure that no promising target genes will be neglected.

Lee McGuire *Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA.*
lmcguire@broadinstitute.org

Revive China's green GDP programme

In a potentially big step towards achieving its target of sustainable growth by 2020, China's government is developing a green measure of gross domestic product (GDP). We suggest that the country's upcoming audit of its natural-resource assets would provide an ideal opportunity to launch this ‘green GDP’, which factors the environmental costs of economic growth into the

conventional GDP.

The government is recognizing that economic growth comes at too high a price. The cost of China's pollution damage roughly quadrupled from 2004–13, and has accounted for up to 3% of annual GDP over the past decade. Each year, there are 350,000 to 500,000 premature deaths from particulates in cities (Z. Chen *et al. Lancet* **382**, 1959–1960; 2013). Indeed, the health cost of air pollution amounts to one-third of total environmental costs.

Although China's original green GDP programme of 2006 was shelved within a year, studies on a green GDP index have never stopped. In the push for green development, faster economic growth is no longer the priority. And under China's latest Five-Year Plan (see *Nature* **531**, 425–426; 2016), local governments are now accountable for environmental quality and ecological conservation.

Jinnan Wang* *Chinese Academy for Environmental Planning, Beijing, China.*

wangjn@caep.org.cn

**On behalf of 7 correspondents (see go.nature.com/rsebg9 for full list).*

Bee-hawking hornet already in line of fire

We agree with Frederico Santarém and colleagues that public campaigns will help to control the invasive Asian hornet *Vespa velutina* (see *Nature* **532**, 177; 2016). However, this bee-hawking hornet has been on Europe's risk-assessment list for invasive alien species and targeted for action since June 2015 (see go.nature.com/gigftz). It has also been intensively researched since 2008 under the European Agricultural Guarantee Fund's apiculture programme.

The only way found so far to contain the *V. velutina* invasion is to destroy colonies as soon as nests are spotted (see J. R. Beggs *et al. BioControl* **56**, 505–526; 2011). Public awareness and collaboration are crucial to help

detect these nests in tree crowns.

The hornet's real threat is to pollinators, not to humans (see, for example, C. Villemant *et al. Biol. Conserv.* **144**, 2142–2150; 2011). EU legislation aims to coordinate a plan for invasion control, which also depends on a greater willingness among European researchers to work together.

Quentin Rome, Claire Villemant *Institut de Systématique, Évolution, Biodiversité (ISYEB), UMR 7205 – CNRS, MNHN, UPMC, EPHE, Sorbonne Universités, Paris, France.*
rome@mnhn.fr

Industry parks limit circular economy

We suggest that China's proposed circular economy should cover the entire life cycle of products and not just focus on industrial parks (see J. A. Mathews and H. Tan *Nature* **531**, 440–442; 2016).

Consumer-waste recycling, for example, should also be part of the circular economy. The delivery of online orders in China last year accounted for some 8 billion plastic bags, 10 billion boxes and 17 billion metres of adhesive tape, yet most of the retailers and companies responsible have no recycling arrangements (see go.nature.com/pv2omq; in Chinese).

Industrial parks designed for a circular economy can have serious limitations, because the interdependence of manufacturers creates a vulnerability. For example, if any one of them closes down or switches to other products, the whole production chain can collapse.

Moreover, these parks cannot be built everywhere. Site selection depends on local manufacturing priorities and on that area's environmental, social and technological conditions. Transforming conventional industrial parks and zones to circular-economy parks has also been problematic because of poor

planning, design and supervision.

Government-controlled circular-economy projects need to be made publicly accountable to safeguard against financial corruption and to ensure transparent oversight. Disorderly operation, enforcement or supervision in the recycling of pollutants or hazardous materials, for example, can lead to disasters such as last year's huge chemical explosion at Tianjin (see Z. Tang *et al. Nature* **525**, 455; 2015).

Xin Miao *Harbin Institute of Technology, Harbin, China.*

Yanhong Tang *Northeast Agricultural University, Harbin, China.*
xin.miao@aliyun.com

Ukraine should cut back nuclear power

Thirty years on from the Chernobyl nuclear disaster, the Ukrainian government has increased the contribution of nuclear power to the nation's total energy balance. In 1991, it was 8%; by 2014, it had risen to 22% (go.nature.com/s4qgjk).

In my view, Ukraine should be following Lithuania's lead. Lithuania has ceased to depend on nuclear power, substituting renewable energy sources (mostly biofuels) for its former 26% nuclear-power contribution in 1991. Renewables are likewise thriving in Latvia (39%) and Estonia (27%) (go.nature.com/z3ibww); Georgia (31%; go.nature.com/5ihqsg); and Kyrgyzstan (28%; go.nature.com/slcnsu).

Ukraine lags way behind because of the funding deficit for new green technologies, with renewables accounting for just 2.6%. This is despite the country's favourable conditions for green energy, including wind, solar and hydropower. This untapped potential could be swiftly realized with appropriate financial and legislative support.

Alexander Gorobets *Sevastopol, Crimea.*
alex-gorobets@mail.ru

Walter Kohn

(1923–2016)

Condensed-matter physicist who revolutionized quantum chemistry.

Walter Kohn's profound questioning of what the arrangement of electrons can tell us about a material's character led to density functional theory. The theory, which predicts electron energies, became a basic tool in efforts to compute the properties of materials and the outcomes of chemical reactions. Some say that it revolutionized quantum chemistry, the application of quantum mechanics to the study of molecules.

Kohn, who died on 19 April, was born in Vienna in 1923. In 1939, not long after the annexation of Austria by Nazi Germany, Kohn's parents sent him to England on a convoy of the *Kindertransport*, an operation to rescue Jewish children from Europe before the outbreak of the war. His mother and father were later killed at Auschwitz.

In 1940, as a holder of a German passport, Kohn was shipped to the first of what would be a series of internment camps in Canada. Once free to leave, he began studies at the University of Toronto, where he earned a bachelor's degree in mathematics and physics and master's degree in applied mathematics. In 1948, he completed a PhD in nuclear physics at Harvard University in Cambridge, Massachusetts; his supervisor was the Nobel-prizewinning theoretical physicist Julian Schwinger.

In 1950, after a short stint of postdoctoral work, Kohn took a professorship at the Carnegie Institute of Technology in Pittsburgh, Pennsylvania (now Carnegie Mellon University). A decade later, he joined the physics department at the University of California, San Diego, where he worked for nearly 20 years before becoming the founding director of the Institute for Theoretical Physics at the University of California, Santa Barbara (now the Kavli Institute).

A condensed-matter system, from a single atom to a living organism, is composed of nuclei and electrons. The electrons roam in an energy landscape provided by the nuclei, and each electron is influenced by the others. The electrical charges of any pair of electrons in the same energy landscape interact, and no electron can exist in the same state as another in the same energy landscape (the Pauli exclusion principle).

In the 1950s and 1960s, physicists were using two different approaches to compute the energy states of electrons in a material. In both approaches, the energy landscape was thought to be key to the prediction of



the properties of a system, including the distribution of electron density. The density functional theory swapped the cause and effect roles between the energy landscape and the electron-density distribution. This paved a way to compute the properties of functional importance to technologies and to life, such as electronics and photosynthesis.

Around 1960, Kohn started to examine the change that occurs to the spatial distribution of the electron density when an impurity is added to a metal. For a positively charged impurity, the electrons pile up around it as expected. They also exhibit a wave-like distribution (Friedel oscillations), which reflects a quantum property of the electrons. This quantum feature led Kohn to examine the possibility that the electron density contained the key to other properties. In 1964, while on sabbatical in Paris, he established with Pierre Hohenberg, a postdoc at the École Normale Supérieure, the Hohenberg–Kohn density theorem. This stated that the electron-density distribution (not the energy landscape) determines the properties of a many-electron system.

Returning to San Diego, Kohn prompted a postdoc, David Mermin, to generalize the theorem so that it could be applied to all temperatures. In 1965, he established with another postdoc (me) a way to use density functional theory to compute the properties of materials.

Kohn's PhD student at San Diego, Philip Tong, was the first to apply density functional theory to infer the electron energies of atoms of noble gases and of the sodium lattice. With his postdoc, Norton Lang, Kohn applied the theory to calculate properties of metal surfaces in the early 1970s. Kohn and Lang won the Davisson–Germer prize in 1977 for their contribution to surface physics. For his work on density functional theory, Kohn shared the 1998 Nobel Prize in Chemistry.

For several years after the Hohenberg–Kohn theorem was published, theoretical chemists raised objections — almost unanimously — to the central role of the electron-density distribution. They could prove that a more general property known as the density matrix was the fount of all electronic properties. They thought that the electron-density distribution, which was only a component of this matrix, could not offer the same predictive power. In the end, people were persuaded by the simplicity of the proof of the theorem, and by the efforts of numerous researchers who showed its usefulness.

Walter was meticulous in his research — but in sports he was adventurous. In 1996, he wrecked his shoulder skiing the day before a widely anticipated talk on density functional theory at an annual meeting of the American Physical Society, and asked me to speak in his stead. He said that he had used a mogul to launch a jump, recalling the ski jumps he had made as a child in Austria. On another occasion, he took his eldest daughter, Marilyn, and me sailing beyond the surf at La Jolla Shores beach in California on a windy day. The boat capsized, and as we pushed it back towards the beach the surf ripped it from our hands.

Walter cared deeply about social issues. At San Diego, he promoted the Judaic studies programme. He was also a vocal critic of the University of California's association with the national weapons laboratories in Los Alamos, New Mexico, and in Livermore, California. And he was proud of producing a documentary film promoting solar energy.

Walter was an admired mentor and colleague, and will be missed by the many who came within his orbit. ■

Lu J. Sham is distinguished professor emeritus of physics at University of California, San Diego, San Diego, California, USA.
e-mail: lsham@ucsd.edu

KARL SCHOENDOERFER/REX/SHUTTERSTOCK

PLANETARY SCIENCE

Pluto's polygons explained

The Sputnik Planum basin of Pluto contains a sheet of nitrogen ice, the surface of which is divided into irregular polygons tens of kilometres across. Two studies reveal that vigorous convection causes these polygons. [SEE LETTERS P.79 & 82](#)

ANDREW J. DOMBARD & SEAN O'HARA

Imagine sunrise on a frozen plane. Shadows withdraw as the Sun climbs above distant mountains that rise from below the horizon. The ice sheet itself is largely featureless, with a difference in elevation of only some tens of metres over distances of many tens of kilometres — nothing slows the shadows' retreat. The sky overhead remains black, and it stays chilly, only about 35 degrees above absolute zero. This is morning on Sputnik Planum, Pluto.

The fly-by of Pluto (and its satellites) by NASA's New Horizons space probe in July 2015 revealed a spectacular world unlike any yet seen¹. The informally named Sputnik Planum is of special interest, a bright, flat-floored basin around 1,200 km in diameter that is filled mainly with nitrogen ice (Fig. 1). High-resolution images¹ show a surface separated into polygonal cells 10–40 km in diameter, pockmarked by pits and fed by flowing nitrogen glaciers from the surrounding highlands.

In this issue, Trowbridge *et al.*² (page 79) and McKinnon *et al.*³ (page 82) investigate this polygonal terrain and conclude that it is continually and quickly resurfaced by convection, making it one of the youngest surfaces in the Solar System. Pluto therefore joins Europa, Enceladus, Titan and Triton as a small and icy but geologically dynamic body of the outer Solar System — a far cry from the cold, dead worlds one might expect so far from the Sun.

The nitrogen ice identified by New Horizons is a structurally weak solid with a very low melting point (63 kelvin), and should flow viscously even at Pluto's low temperatures⁴. As in other planetary bodies, Pluto's interior is warmer than its surface because it is heated by the decay of long-lived radiogenic isotopes in the rocky component. How this heat escapes through Sputnik Planum has consequences for its surface geology. For a layer of weak nitrogen ice at least 0.5–1 km thick, the most efficient heat-transfer mechanism is convection.

Because the material is heated from the

bottom, the heat will cause localized thermal expansion, making the heated material less dense than the rest of the overlying ice. In convection, the less-dense material is buoyant and will rise, carrying its heat content towards the surface, where it cools and then sinks. Viscous drag resists this buoyancy-driven movement, and convection can occur only if the buoyancy overwhelms the viscous resistance.

This competition can be quantified. The ratio of buoyant to viscous forces in a layer defines a dimensionless parameter known as the Rayleigh number. If the Rayleigh number is greater than a critical value, then the material convects. Both Trowbridge *et al.* and McKinnon *et al.* found that the Rayleigh number of the polygonal terrain is several orders of magnitude greater than the critical value. Their results indicate that the nitrogen ice is vigorously convecting and that the cellular patterns are the tops of convection cells.

In addition, both groups report that the convective flow speeds are in the range of centimetres per year, meaning that the surface

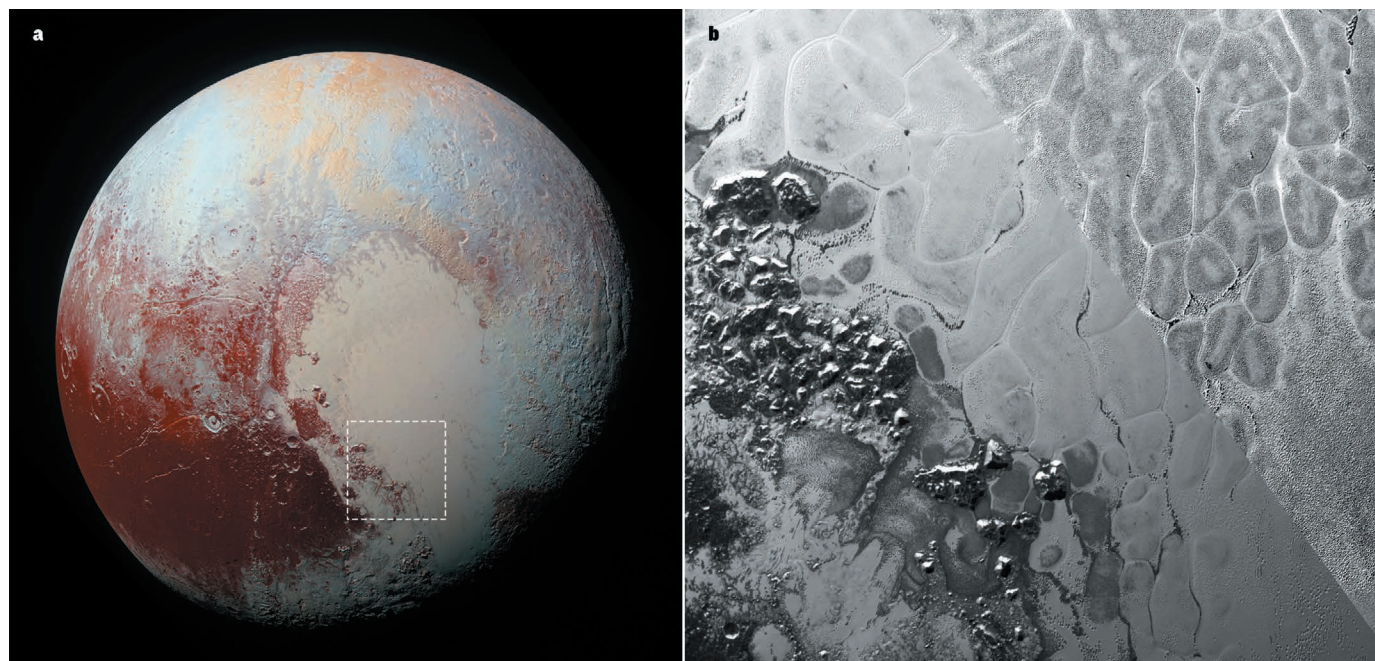


Figure 1 | Sputnik Planum. **a**, One of Pluto's youngest terrains, informally known as Sputnik Planum, is a nitrogen ice sheet (visible here as the large pale expanse) that fills a topographic basin. **b**, This composite image showing a closer view of the nitrogen ice reveals irregular polygons about 10–40 kilometres in diameter (upper part of image). Two papers^{2,3} report that the polygonal terrain is caused by convection in the nitrogen ice.

NASA/JOHNS HOPKINS UNIV. APPL. PHYSICS LAB./
SOUTHWEST RES. INST.

turns over in about 500,000 to 1 million years. This rapid resurfacing explains the lack of impact craters on the ice sheet. (In general, the older a planetary surface, the more impact craters will have formed.)

Although the two papers report the same primary result, they differ in their conclusions about the convective regime, which determines the width-to-depth aspect ratio of the convection cells and hence the thickness of the nitrogen-ice layer. Trowbridge *et al.* argue that variations in the viscosity of the ice due to differences in stress and temperature across the layer are small enough that convection occurs in the Rayleigh–Bénard regime, which is characterized by the formation of cells that have widths similar to their depths⁵. The cell size of 10–40 km thus implies a layer thickness of at least 10 km.

By contrast, McKinnon *et al.* argue that the temperature dependence of the nitrogen ice causes ‘sluggish lid’ convection, in which the viscosity is higher at the cooler surface than in the interior⁵. As the name suggests, this yields a slower-moving surface layer and cells that

are much wider than they are deep, making the depth of the layer 3–6 km. The authors support this conclusion with numerical modelling that reproduces convection cells with sizes and surface topography that are consistent with observations.

The layer thickness has important implications for Pluto’s geological history. On the basis of the shape and ellipticity of the basin that holds Sputnik Planum, McKinnon and colleagues note that it is most probably an ancient impact crater³. From scaling of other examples in the Solar System, it is known that an impact basin of this size can easily accommodate the depth of nitrogen ice estimated by McKinnon *et al.*, but not the depth estimated by Trowbridge and colleagues. Their deeper prediction requires a more complicated explanation of basin formation and evolution. Perhaps the weight of the nitrogen ice caused the basin to subside, for example.

Both papers report that the quantity of ice in the basin is equivalent to a global layer several hundred metres in depth, commensurate with Pluto’s total budget of nitrogen. But

neither satisfactorily addresses how so much of the nitrogen budget could have collected there — was it for climatological reasons, as Trowbridge and co-workers speculate, or for glaciological reasons, as McKinnon *et al.* suggest? Clearly, this localization of nitrogen was a major event in Pluto’s evolution that needs to be explored. Fortunately, New Horizons continues to transmit data from its Pluto encounter back to Earth. It is to be hoped that these two papers will be the first step towards a deeper understanding of this distant world. ■

Andrew J. Dombard and Sean O’Hara are in the Department of Earth and Environmental Sciences, University of Illinois at Chicago, Chicago, Illinois 60607-7059, USA.

1. Stern, S. A. *et al.* *Science* **350**, aad1815 (2015).
2. Trowbridge, A. J., Melosh, H. J., Steckloff, J. K. & Freed, A. M. *Nature* **534**, 79–81 (2016).
3. McKinnon, W. B. *et al.* *Nature* **534**, 82–85 (2016).
4. Yamashita, Y., Kato, M. & Arakawa, M. *Icarus* **207**, 972–977 (2010).
5. Schubert, G., Turcotte, D. & Olson, P. *Mantle Convection in the Earth and Planets* (Cambridge Univ. Press, 2001).

MICROBIOLOGY

Pumping persisters

The finding that antibiotics are pumped out of drug-tolerant bacterial cells by the TolC protein complex provides insight into how some cells, known as persisters, survive in the face of antibiotic treatments.

KENN GERDES & SZABOLCS SEMSEY

In bacterial persistence, a small fraction of an antibiotic-sensitive cell population has switched to a slow-growing or dormant state, and is drug tolerant^{1,2}. This differs from antibiotic resistance in that regrowth of a persistent population results in the same percentage of drug-sensitive cells as before. Persistence has been interpreted as a bet-hedging strategy that increases the survival rates of bacterial populations^{3,4}, and is medically relevant because it might sustain recurrent and chronic infections. Writing in *Molecular Cell*, Pu *et al.*⁵ challenge the widespread view that persistence is a passive state. The authors demonstrate that persister cells use an energy-dependent efflux pump protein called TolC to actively reduce the intracellular accumulation of antibiotic — a finding that might have both fundamental and therapeutic relevance.

Pu and colleagues isolated persisters and labelled them with a fluorescent antibiotic called BOCILLIN, which is derived from penicillin. They observed the cells using microscopy and found that the antibiotic could penetrate persister cells. However, the average antibiotic concentration in the persisters

was about 20% of that in the drug-sensitive population.

TolC is the outer-membrane component of a family of efflux pumps that can move small molecules out of the cell from both the cytoplasm and the periplasmic space between the

inner and outer bacterial membranes. Using sophisticated microfluidics combined with fluorescence microscopy, Pu *et al.* showed that TolC is responsible for the rapid export of BOCILLIN from persisters (Fig. 1). An alternative explanation could be that less of the antibiotic is taken up into cells in the first place, but the authors found that lower membrane permeability owing to depletion of porin proteins only slightly decreased BOCILLIN uptake. These observations raised the possibility that increased TolC levels contribute to drug tolerance in persisters.

Next, an analysis of cells in which TolC was labelled with a fluorescent dye called FLA_{SH} revealed that persisters do have higher TolC levels than the drug-sensitive subpopulation.

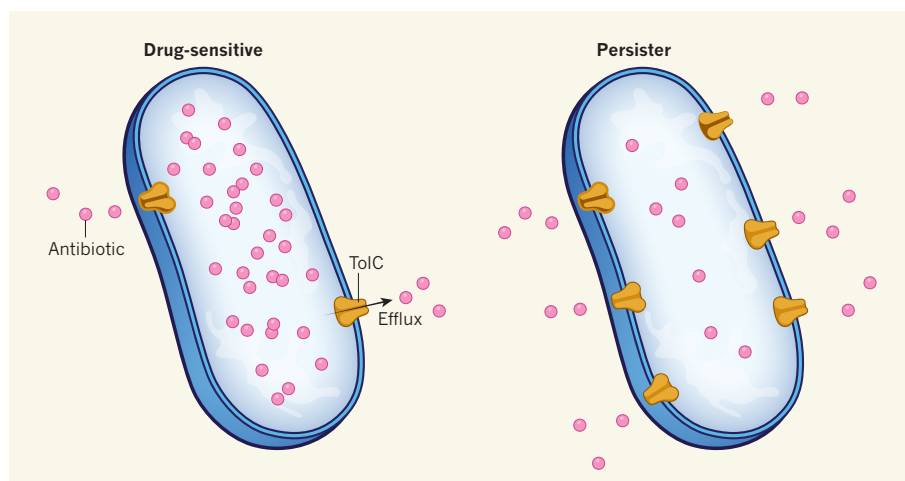


Figure 1 | A subpopulation in efflux. Persistence is a phenomenon whereby a small fraction of cells in a bacterial population survive antibiotic treatment. Pu *et al.*⁵ demonstrate that persister cells upregulate production of the TolC protein relative to drug-sensitive cells. TolC is part of a membrane-spanning efflux pump that transports antibiotic out of the cell, thus promoting survival.

Moreover, when the authors isolated the subpopulation of cells with relatively high TolC levels, they found that this fraction contained nearly 20 times more persisters than the rest of the population. Thus, there is a clear correlation between a high level of TolC and persistence. The researchers then tracked cells using the FAsH-labelled TolC: these experiments suggested that most persister cells emerged from a subpopulation that had increased levels of TolC even before treatment with the antibiotic. This important result warrants further study, and raises the question of whether the molecular mechanism that underlies the drug-independent variation of TolC is separate from, or an integral component of, other pathways that are already known to regulate stochastically induced persistence.

The current study leaves little doubt that TolC is involved in persistence. Most convincingly, perhaps, Pu and colleagues showed that deletion of the *tolC* gene or inhibition of TolC with a chemical compound drastically reduced the level of persisters. Because TolC is an outer-membrane protein, such inhibitors can readily access the protein. These observations raise the question of whether it might, in the future, be possible to develop therapeutic co-drugs that increase the efficacy of conventional antibiotics. These could be particularly useful for treating chronic and recurrent infections.

Many other genes have previously been implicated in bacterial persistence, including toxin–antitoxin (TA) genes. Most type II TA genes encode inhibitors of translation — their expression might therefore contribute to the dormancy of persisters^{2,3}. Indeed, deletion of several type II TA genes significantly reduces persistence in the bacterium *Escherichia coli*⁶ and in a subspecies of *Salmonella enterica*⁷. The small membrane proteins encoded by type I TA genes can also induce persistence, by depolarizing the membrane, thereby reducing cellular levels of the energy-carrying molecule ATP and thus contributing to dormancy⁸.

Expression of both type I and II TAs is induced stochastically by the signalling molecules guanosine tetra- and pentaphosphate, and so the two classes might contribute synergistically to dormancy by reducing ATP levels and protein synthesis, respectively. How could the stochastic variation of TolC levels observed by Pu *et al.* fit into the regulatory scheme that controls type I and II TAs? Expression of the *tolC* gene is regulated by several transcriptional activators that respond to chemical compounds, including antibiotics, but if expression is also induced stochastically before chemical stress, TolC might act in concert with type I and II TAs to increase the drug tolerance of persisters. This would be the first example of an active mechanism contributing to stochastically induced multiple-drug tolerance. More research is required to resolve this exciting, outstanding question. ■

Kenn Gerdes and Szabolcs Semsey are in the Department of Biology, University of Copenhagen, DK-2200 Copenhagen, Denmark.
e-mail: kgerdes@bio.ku.dk

1. Bigger, J. W. *Lancet* **ii**, 497–500 (1944).
2. Balaban, N. Q., Merrin, J., Chait, R., Kowalik, L. & Leibler, S. *Science* **305**, 1622–1625 (2004).
3. Maisonneuve, E., Castro-Camargo, M. & Gerdes, K.

- Cell* **154**, 1140–1150 (2013).
4. Veening, J.-W., Smits, W. K. & Kuipers, O. P. *Annu. Rev. Microbiol.* **62**, 193–210 (2008).
5. Pu, Y. *et al. Mol. Cell* **62**, 284–294 (2016).
6. Maisonneuve, E., Shakespeare, L. J., Jørgensen, M. G. & Gerdes, K. *Proc. Natl Acad. Sci. USA* **108**, 13206–13211 (2011).
7. Helaine, S. *et al. Science* **343**, 204–208 (2014).
8. Verstraeten, N. *et al. Mol. Cell* **59**, 9–21 (2015).

This article was published online on 25 May 2016.

STRUCTURAL BIOLOGY

A photo shoot of plant photosystem II

In photosynthesis, the plant photosystem II uses the energy in sunlight to oxidize water. The high-resolution structure of this crucial supercomplex has now been obtained using cryo-electron microscopy. [SEE ARTICLE P.69](#)

ROBERTA CROCE & PENGQI XU

Photosystem II is the enzyme complex that produces the oxygen we breathe. It is at the heart of the photosynthesis process, and uses the energy of the Sun to extract from water the electrons and protons that are needed to produce food and fuel. On page 69 of this issue, Wei *et al.*¹ report the structure of spinach photosystem II — a 1.1-megadalton dimeric complex in which each monomer is composed of 25 proteins and 133 pigment molecules. This structure provides a plethora of information to aid our understanding of the molecular mechanisms by which light is converted into chemical energy.

Photosystem II (PSII) is a membrane-embedded modular assembly of pigment–protein complexes and is composed of two main parts, the core and the outer antenna. The core contains the reaction centre in which energy is used to drive photochemistry. It has an evolutionarily highly conserved protein composition in all the organisms that perform oxygen-generating photosynthesis².

Wei and colleagues' structure shows that both the protein and the pigment organization of the plant PSII core in the membrane region are almost identical to those of the core of the previously reported³ structure of cyanobacterial PSII; this indicates that the complex was optimized long ago and has not changed since. Only the peripheral membrane proteins that surround the water-splitting catalyst in the core are organized differently in plants and cyanobacteria. This observation is intriguing, because these peripheral proteins are needed for oxidizing water⁴, and their organizational differences from their cyanobacterial counterparts might inform how nature has optimized this essential reaction.

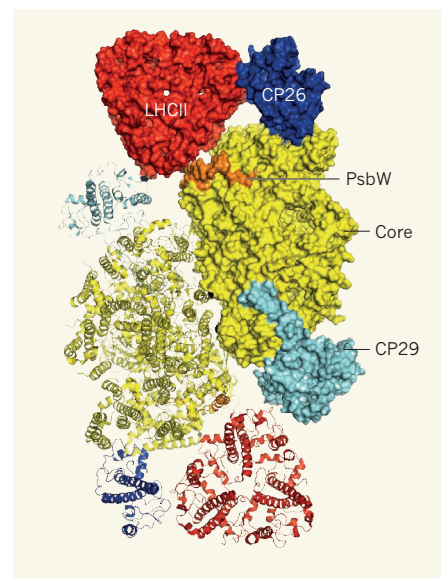


Figure 1 | Structure of the plant photosystem II supercomplex. Wei *et al.*¹ report the structure of spinach photosystem II as a dimeric supercomplex. (Here, one of the monomers is represented as ribbons and the other as its surface area.) Of the two main parts of each monomer, the core complex contains the reaction centre (not shown), and the peripheral light-harvesting complexes (LHCII, CP29 and CP26) supply excitation energy to the reaction centre. PsbW is a core subunit typical of plants, which mediates the association of LHCII with the core.

In contrast to the core, the outer antenna differs greatly between photosynthetic organisms. Its role is to increase the core's capacity to harvest light, overcoming the fact that light is a dilute form of energy (one molecule of the pigment chlorophyll absorbs only a few photons per second even on a bright, sunny day). The outer antenna is shaped by the host

organism's adaptation to its ecological niche, where light quantity and quality vary, and thus it is tailor-made⁵. In vascular plants (including spinach), the outer antenna is composed of light-harvesting complexes (LHCs). These are pigment–protein complexes that absorb light and transfer part of the corresponding energy to the reaction centre.

In the present structure, each monomer of the PSII supercomplex is composed of a core, one LHC trimer (called LHCII; ref. 6) and one monomer of each of two minor LHCs, CP29 and CP26. Wei and colleagues' work offers the first structures of the plant PSII core, CP26 and the complete CP29. It also shows the position of the core subunit PsbW and the molecular details of the connection between the antenna and the core (Fig. 1). Notably, the long amino-terminal region of CP29 extends all the way over the CP47 subunit of the core to interact with the D1 protein of the reaction centre. This organization provides a structural basis for the observation⁷ that, in plants, LHCs are required for the connectivity in the core.

A substantial problem in obtaining structural information about plant PSII has been its instability, which is a direct consequence of its functional behaviour. Not only must PSII adjust its antenna size in response to natural lighting conditions, but it must also repair its reaction centre; splitting water using sunlight is a risky business that can lead to the generation of reactive oxygen species, which damage the system. To repair the damage, PSII undergoes regular 'pit stops', during which it is disassembled and reassembled after substitution of the damaged part⁸. Cryo-electron microscopy was therefore crucial in solving this structure, because it allowed Wei *et al.* to select the intact particles from the ensemble and reach an amazing 3.2-ångström resolution.

Although modularity is required to allow repair, a good connection between the subunits is equally important for efficient energy transfer from the antenna to the core. The absorption of a photon promotes chlorophyll to an excited state, but this state is unstable and the chlorophyll relaxes to its initial (ground) state within a few nanoseconds. Once back to the ground state, the energy is lost. Consequently, time is limited for using the energy, which must then be rapidly transferred to the reaction centre.

Earlier work showed⁷ that in the plant PSII supercomplex, it takes 140 picoseconds (1 ps is 10⁻¹² s) from the absorption of a photon by a chlorophyll molecule to the charge separation in the reaction centre, such that more than 90% of the photons absorbed produce an electron. How fast this energy is transferred from the antenna to the reaction centre depends on the distance between the pigments, their relative orientation and their energy. These factors are all dictated by the proteins, which act as smart matrices that organize the pigments.

The LHCs contain two types of chlorophyll (*a* and *b*) that are chemically similar but

energetically different. Chlorophyll *a* absorbs lower-energy photons than chlorophyll *b*. Because energy preferentially migrates downstream, chlorophyll *b* rapidly transfers it to chlorophyll *a*; the energy is then transferred to the reaction centre mainly by chlorophyll *a*.

Although detailed computational modelling based on the new structure is needed for a quantitative understanding of the excitation-energy transfer, visual inspection of how chlorophylls are organized in the supercomplex already provides qualitative indications. Intriguingly, the interface between the LHCs is occupied by chlorophyll *b* molecules located relatively far from each other, suggesting that there is little (if any) transfer of energy between the LHCs present in this supercomplex. Instead, all LHCs seem to transfer their collected energy directly to the core.

This organization may seem at odds with the required efficiency, because more energy-transfer pathways normally result in higher efficiency⁹. In cells, however, PSII is in contact with other LHCs, the number of which varies under different conditions and which are not present in the isolated supercomplex¹⁰. The presence of chlorophyll *a* at the periphery of the supercomplex can help to transfer energy from such additional LHCs to the core.

Clearly, Wei *et al.* have provided us with

a wonderful structure. The ball is now in the court of spectroscopists and theoreticians to use this structure to obtain a detailed understanding of the functionality of the system. ■

Roberta Croce and Pengqi Xu are in the Biophysics of Photosynthesis Group, Department of Physics and Astronomy, Faculty of Science, VU University Amsterdam, 1081 Amsterdam, the Netherlands. e-mail: r.croce@vu.nl

1. Wei, X. *et al.* *Nature* **534**, 69–74 (2016).
2. Hohmann-Marriott, M. F. & Blankenship, R. E. *Annu. Rev. Plant Biol.* **62**, 515–548 (2011).
3. Umena, Y., Kawakami, K., Shen, J.-R. & Kamiya, N. *Nature* **473**, 55–60 (2011).
4. Ifuku, K. *Biosci. Biotechnol. Biochem.* **79**, 1223–1231 (2015).
5. Croce, R. & van Amerongen, H. *Nature Chem. Biol.* **10**, 492–501 (2014).
6. Liu, Z. *et al.* *Nature* **428**, 287–292 (2004).
7. Caffarri, S., Broess, K., Croce, R. & van Amerongen, H. *Biophys. J.* **100**, 2094–2103 (2011).
8. Järvi, S., Suorsa, M. & Aro, E.-M. *Biochim. Biophys. Acta Bioenerget.* **1847**, 900–909 (2015).
9. Chmeliov, J., Trinkunas, G., van Amerongen, H. & Valkunas, L. *J. Am. Chem. Soc.* **136**, 8963–8972 (2014).
10. Dekker, J. P. & Boekema, E. J. *Biochim. Biophys. Acta* **1706**, 12–39 (2005).

This article was published online on 18 May 2016.

ARCHAEOLOGY

Neanderthals built underground

The finding of 175,000-year-old structures deep inside a cave in France suggests that Neanderthals ventured underground and were responsible for some of the earliest constructions made by hominins. SEE LETTER P.111

MARIE SORESSI

Building is a frequent by-product of human activity, and some ancient constructions remain majestic to this day. However, all too often, constructions made by mobile populations do not preserve well. Evidence for structures made by prehistoric hunter–gatherers are scarce and usually consist only of an area of finds with intriguing spatial distributions, which may be associated with a fireplace. On page 111 of this issue, Jaubert *et al.*¹ report the discovery of circular structures made of broken stalagmites deep inside a cave in southwest France. The structures are up to 40 centimetres high and 6.7 metres wide, and direct radiometric dating shows that they are at least 175,000 years old. Because Neanderthals were the only hominin group present in western Europe at that time, the discovery provides the first directly dated

evidence for Neanderthals' construction abilities. It also shows that Neanderthals explored underground.

Neanderthals lived in Eurasia from around 400,000 to 40,000 years ago, at which point anatomically modern humans settled in. Investigation of the archaeological record from the Late Pleistocene epoch — which spanned from 126,000 to 11,700 years ago — has provided robust data on the behaviour of ancient hominins and allowed a comparison of the activities of Neanderthals and early modern humans. This comparative approach has been regularly used to elaborate on the reasons for Neanderthals' demise and the success of early modern humans.

However, given a lack of direct evidence, there has been little discussion of the constructional abilities of Neanderthals. It is known that great apes, birds and other animals build



Figure 1 | Ancient structures. Circular structures made from broken stalagmites, found in Bruniquel Cave in southwest France by Jaubert *et al.*¹, are thought to have been made by Neanderthals around 175,000 years ago.

elaborate nests (the bowerbird is a famous example), and the archaeological record contains examples of constructions made by anatomically modern humans about 20,000 years ago, such as collapsed, rounded 'ruins' made from mammoth bones or deer antlers². Yet only a few structures interpreted as post-holes or isolated elements of dry stone walls have been tentatively attributed to Neanderthals. Furthermore, differential distributions of finds inside and outside potential Neanderthal constructions have rarely been documented, and even then not always convincingly².

Jaubert *et al.* report accumulations of almost 400 stalagmites and stalagmite fragments stacked into several structures, including two that have a semicircular shape, some 300 m from the entrance of Bruniquel Cave (Fig. 1). One semicircular structure, which is more than 6.7 m wide, comprises a 'wall' made of up to four superimposed layers of stalagmite fragments about 30 cm in length, with smaller elements stuck obliquely in between. Reddening, blackening and cracking of many stalagmites suggest that the structures have been heated by small fires. The authors also recovered a 6.7-cm fragment of heated bone from within one of the smaller structures, close to reddened and blackened stalagmites. This find, together with measurement of the magnetic anomalies in the rock above and around the structures, supports the idea that the structures were heated.

The researchers used molecular and atomic spectrometry to investigate two other probable residues of heated bones, one found in a 2-m-wide structure and the other

forming part of a concentration of similarly blackened material discovered on the ground and interpreted as a hearth. Seven stalagmites from the two largest structures were dated using uranium-series measurements; by dating the calcite that had grown before and after the fragments were broken, the researchers could constrain the date at which the stalagmites were used in construction. The calcite covering the 6.7-cm-long bone and forming the flowstone (a sheet-like calcite deposit) on the floor of the largest structure was also dated.

Altogether, the authors dated 18 samples from the area containing the structures, which show that the structures are around 176,500 years old (with a confidence interval of 2,100 years). That period is known to have had relatively warm and humid phases, which is consistent with the calcite deposition observed. The signal of oxygen and carbon isotopes reported from the stalagmites is also consistent with the atmospheric conditions known for that time.

The inner organization and the size of the structures do not fit with what is documented for the nests of cave bears, discounting that possibility for their construction. Thus, these structures are the oldest directly dated constructions attributed to Neanderthals, and the first ones for which we can be confident of that attribution. Furthermore, no charred materials have been found outside the structure, and no red- or black-coloured material was observed on the cave ceiling above the structure: these details support the idea that the colorations are indicative of heating *in situ* and were not transported between or onto the stalagmites by natural processes.

Jaubert *et al.* discuss the social organization that would have been needed to manufacture such structures, and compare this with what is known for modern humans from the same era. They conclude that their discovery indicates that Neanderthals exhibited more-complex social behaviour than was previously thought, and suggests that these hominins used the underground environment. Only further discovery of underground structures will help to establish whether these structures were opportunistic ones relating to an accidental underground visit, or whether they were part of regular and planned Neanderthal activities.

These structures are among the best-preserved constructions known for the whole of the Pleistocene epoch, probably because they were sealed by calcite very soon after they were erected. When the best evidence is found in the best-preserved context, it serves as a reminder for archaeologists of how much we depend on preservation. The fact that some of the art of the period is also often found deep inside caves has been alternatively interpreted as a testimony of the preservation provided by the cave environment³ or as a result of spiritual preoccupations — the underground being a special place⁴. Perhaps we need to further consider the idea that the fuzziness of the Neanderthal record is due to a lack of preservation. Given that we often discuss archaeological findings in a comparative framework that contrasts Neanderthals (which disappeared) with early modern humans (who were obviously successful), we may also wonder how this framework is biased by Western thought. European culture is known for having emphasized what may be 'uniquely human' and may separate 'us' from other animals.

Comparing hominins across a large chunk of time is necessary and useful. However, an increased focus on reconstructing the historical context of past behavioural and technological innovations may be key to further understanding these different populations. The structures discovered by Jaubert *et al.* are a good example of how reconstructing ancient history may benefit from not only broad-scale comparisons of evolution over time but also detailed analysis of specific areas at specific time points. ■

Marie Soressi is in the Faculty of Archaeology, Leiden University, 2300 RA Leiden, the Netherlands.

e-mail: m.a.soressi@arch.leidenuniv.nl

1. Jaubert, J. *et al.* *Nature* **534**, 111–114 (2016).
2. Klein, R. G. *The Human Career: Human Biological and Cultural Origins* 3rd edn (Univ. Chicago Press, 2009).
3. Guthrie, R. D. *The Nature of Paleolithic Art* (Univ. Chicago Press, 2005).
4. Clottes, J. & Lewis-Williams, J. D. *Les Chamanes de la Préhistoire: Transe et Magie dans les Grottes Ornées: Suivi de Après les Chamanes, Polémiques et Réponses* (Seuil, 1996).

This article was published online on 25 May 2016.

GEOPHYSICS

Earth's core problem

Measurements of the electrical resistance and thermal conductivity of iron at extreme pressures and temperatures cast fresh light on controversial numerical simulations of the properties of Earth's outer core. **SEE LETTERS P.95 & 99**

DAVID DOBSON

Earth's core acts like a storage heater, with heat released during crystallization of the inner core that buffers the slow cooling of the planet as it radiates its heat to space. The most obvious expression of this heat transfer is Earth's magnetic field, which is generated by convection in the liquid outer core. But the magnitude of the transfer is controlled by thermal conduction across the boundary between the core and mantle.

In 2012, first-principles numerical simulations^{1,2} indicated that the thermal conductivity of liquid iron in the outer core is so high that this region might act as a pump that pushes heat towards the core–mantle boundary faster than convection can. If, as these controversial studies suggest, the core is losing heat at such a high rate, it means that the magnetic field must work in previously unimagined ways³, and that the solid inner core must be less than a billion years old⁴ — a mere babe in planetary terms. In this issue, Ohta *et al.*⁵ (page 95) and Konôpková *et al.*⁶ (page 99) report studies that experimentally tested the simulations' results using complementary, but distinct, approaches and come to different conclusions.

Both groups use laser-heated diamond–anvil cells to generate the extreme temperatures and pressures of the core–mantle boundary, but that is where the similarity ends. Ohta *et al.* measured the electrical resistance of iron wires, which is closely related to the wires' thermal conductivity (Fig. 1a). To convert the resistivity measurements to a measure of the thermal conductivity of liquid iron in the outer core, the authors fitted their data to a model of resistivity that assumes that resistance approaches a limit at high temperature (a phenomenon called resistivity saturation). This then allowed them to use the Wiedemann–Franz relationship between resistance and thermal conduction in metals to calculate the thermal conductivity. Both of these procedures have good theoretical bases and are well established for low-pressure observations. The observed high electrical conductivities resulted in a predicted outer-core thermal conductivity of around 90 watts per metre per kelvin, which is in reasonable agreement with the 2012 simulations^{1,2}.

By contrast, Konôpková *et al.* directly measured thermal conduction by watching a heat pulse propagate through a solid iron sample after heating with a nanosecond laser

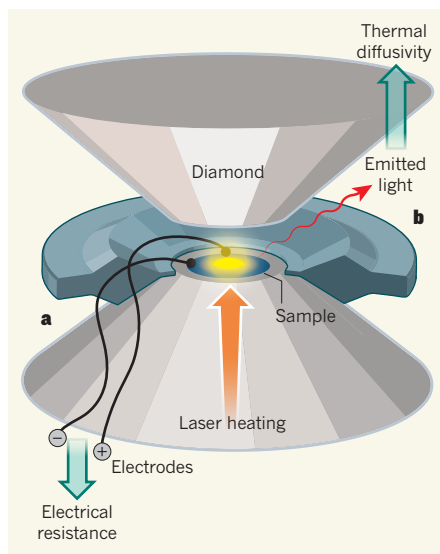


Figure 1 | Measuring the thermal conductivity of iron at Earth's core conditions. In diamond anvil cells, the pressure generated between the tips of diamonds can exceed millions of atmospheres. Lasers can be fired through the diamonds to directly heat a sample of a material to 4,000 kelvin or more. **a**, Ohta *et al.*⁵ connected electrodes to a sample of solid iron and measured its electrical resistance (which is inversely proportional to thermal conductivity in metals) at high temperatures and pressures. **b**, In separate experiments, Konôpková *et al.*⁶ pulsed the laser, and measured the time taken for heat pulses to diffuse through a solid iron sample on the basis of changes in the brightness and wavelength of the light emitted from the sample. This allowed them to measure the thermal rate of diffusion, which is closely related to thermal conductivity.

pulse (Fig. 1b). The time taken for the pulse to pass from the heated side of the sample to the other side, and the amplitude difference of the pulse between the two sides, are functions of the thermal conductivity of the sample, as well as of the surrounding solid medium that transmits pressure from the diamonds to the sample and thermally insulates the sample from the diamonds. After some careful mathematical modelling of the temperature field in the diamond cell, the authors extracted the thermal conductivity of iron from time-resolved changes in the brightness and wavelength of the glow from the white-hot sample. They obtained a thermal conductivity of about $30 \text{ W m}^{-1} \text{ K}^{-1}$, similar to early predictions of outer-core conductivity⁷.

But this leaves us with a conundrum: how to reconcile the high thermal conductivity reported by Ohta and colleagues on the basis of resistance measurements with the low thermal conductivity measured by Konôpková and co-workers. Maybe there were unknown complications with the experiments? For example, the extremely short laser pulses used by Konôpková *et al.* might have caused the sample to partially melt for a short period, which could have gone unnoticed during the experiment. If so, then the melting phase transition would have acted as a thermal buffer (much as the crystallization of the inner core buffers Earth's temperature) and caused an apparent decrease in thermal conductivity. This might explain why the measured thermal conductivities decrease so strongly with temperature, particularly at temperatures approaching the melting temperature.

Or maybe Ohta *et al.* underestimated the heat loss through the electrodes in their experiments, which would mean that the average sample temperature was less than the measured value. This could have made it look as though resistivity was saturating, even if it wasn't. Alternatively, the proportionality constant between electrical resistance and thermal conduction (the Lorenz number) might become strongly temperature dependent at the extreme pressures and temperatures of the experiment — this would point to previously unobserved fundamental physics.

Despite the discrepancy, these two studies are experimental feats, measuring complex physical properties of samples smaller than a pinhead at pressures greater than 1 million atmospheres, and at temperatures above 4,000 K. The fact that the results agree within a factor of three is a remarkable success, but the devil is in the detail. The discrepancy makes a big difference to estimates of when the inner core formed, and hence when Earth generated a stable magnetic field — the inner core could be as little as 700 million years old, about the same age as complex life; or as much as 3 billion years old, about three-quarters of Earth's age. More experimental and theoretical work is needed to resolve the discrepancy and hence to constrain the age of the inner core and the workings of Earth's magnetic field. ■

David Dobson is in the Department of Earth Sciences, University College London, London WC1E 6BT, UK.
e-mail: d.dobson@ucl.ac.uk

1. Pozzo, M., Davies, C., Gubbins, D. & Alfè, D. *Nature* **485**, 355–358 (2012).
2. deKoker, N., Steinle-Neumann, G. & Vlček V. *Proc. Natl Acad. Sci. USA* **109**, 4070–4073 (2012).
3. Buffett, B. *Nature* **485**, 319–320 (2012).
4. Labrosse, S. *Phys. Earth Planet. Inter.* **247**, 36–55 (2015).
5. Ohta, K., Kuwayama, Y., Hirose, K., Shimizu, K. & Ohishi, Y. *Nature* **534**, 95–98 (2016).
6. Konôpková, Z., McWilliams, R. S., Gómez-Pérez, N. & Goncharov, A. F. *Nature* **534**, 99–101 (2016).
7. Stacey, F. D. & Anderson, O. L. *Phys. Earth Planet. Inter.* **124**, 153–162 (2001).

Landscape of somatic mutations in 560 breast cancer whole-genome sequences

Serena Nik-Zainal^{1,2}, Helen Davies¹, Johan Staaf³, Manasa Ramakrishna¹, Dominik Glodzik¹, Xueqing Zou¹, Inigo Martincorena¹, Ludmil B. Alexandrov^{1,4,5}, Sancha Martin¹, David C. Wedge¹, Peter Van Loo^{1,6}, Young Seok Ju¹, Marcel Smid⁷, Arie B. Brinkman⁸, Sandro Morganello⁹, Miriam R. Aure^{10,11}, Ole Christian Lingjærde^{11,12}, Anita Langerød^{10,11}, Markus Ringnér³, Sung-Min Ahn¹³, Sandrine Boyault¹⁴, Jane E. Brock¹⁵, Annegien Broeks¹⁶, Adam Butler¹, Christine Desmedt¹⁷, Luc Dirix¹⁸, Serge Dronov¹, Aquila Fatima¹⁹, John A. Foekens⁷, Moritz Gerstung¹, Gerrit K. J. Hooijer²⁰, Se Jin Jang²¹, David R. Jones¹, Hyung-Yong Kim²², Tari A. King²³, Savitri Krishnamurthy²⁴, Hee Jin Lee²¹, Jeong-Yeon Lee²⁵, Yilong Li¹, Stuart McLaren¹, Andrew Menzies¹, Ville Mustonen¹, Sarah O'Meara¹, Iris Pauporté²⁶, Xavier Pivot²⁷, Colin A. Purdie²⁸, Keiran Raine¹, Kamna Ramakrishnan¹, F. Germán Rodríguez-González⁷, Gilles Romieu²⁹, Anieta M. Sieuwerts⁷, Peter T. Simpson³⁰, Rebecca Shepherd¹, Lucy Stebbings¹, Olafur A. Stefansson³¹, Jon Teague¹, Stefania Tommasi³², Isabelle Treilleux³³, Gert G. Van den Eynden^{18,34}, Peter Vermeulen^{18,34}, Anne Vincent-Salomon³⁵, Lucy Yates¹, Carlos Caldas³⁶, Laura van't Veer¹⁶, Andrew Tutt^{37,38}, Stian Knappskog^{39,40}, Benita Kiat Tee Tan^{41,42}, Jos Jonkers¹⁶, Åke Borg³, Naoto T. Ueno²⁴, Christos Sotiriou¹⁷, Alain Viari^{43,44}, P. Andrew Futreal^{1,45}, Peter J. Campbell¹, Paul N. Span⁴⁶, Steven Van Laere¹⁸, Sunil R. Lakhani^{30,47}, Jorunn E. Eyfjord³¹, Alastair M. Thompson^{28,48}, Ewan Birney⁹, Hendrik G. Stunnenberg⁸, Marc J. van de Vijver²⁰, John W. M. Martens⁷, Anne-Lise Børresen-Dale^{10,11}, Andrea L. Richardson^{15,19}, Gu Kong²², Gilles Thomas⁴⁴ & Michael R. Stratton¹

We analysed whole-genome sequences of 560 breast cancers to advance understanding of the driver mutations conferring clonal advantage and the mutational processes generating somatic mutations. We found that 93 protein-coding cancer genes carried probable driver mutations. Some non-coding regions exhibited high mutation frequencies, but most have distinctive structural features probably causing elevated mutation rates and do not contain driver mutations. Mutational signature analysis was extended to genome rearrangements and revealed twelve base substitution and six rearrangement signatures. Three rearrangement signatures, characterized by tandem duplications or deletions, appear associated with defective homologous-recombination-based DNA repair: one with deficient BRCA1 function, another with deficient BRCA1 or BRCA2 function, the cause of the third is unknown. This analysis of all classes of somatic mutation across exons, introns and intergenic regions highlights the repertoire of cancer genes and mutational processes operating, and progresses towards a comprehensive account of the somatic genetic basis of breast cancer.

The mutational theory of cancer proposes that changes in DNA sequence, termed 'driver' mutations, confer proliferative advantage on a cell, leading to outgrowth of a neoplastic clone¹. Some driver mutations are inherited in the germline, but most arise in

somatic cells during the lifetime of the cancer patient, together with many 'passenger' mutations not implicated in cancer development¹. Multiple mutational processes, including endogenous and exogenous mutagen exposures, aberrant DNA editing, replication errors

¹Wellcome Trust Sanger Institute, Hinxton, Cambridge CB10 1SA, UK. ²East Anglian Medical Genetics Service, Cambridge University Hospitals NHS Foundation Trust, Cambridge CB2 9NB, UK.

³Division of Oncology and Pathology, Department of Clinical Sciences Lund, Lund University, Lund SE-223 81, Sweden. ⁴Theoretical Biology and Biophysics (T-6), Los Alamos National Laboratory, Los Alamos, NM 87545, New Mexico, USA. ⁵Center for Nonlinear Studies, Los Alamos National Laboratory, Los Alamos, New Mexico 87545, USA. ⁶Department of Human Genetics, University of Leuven, B-3000 Leuven, Belgium. ⁷Department of Medical Oncology, Erasmus MC Cancer Institute and Cancer Genomics Netherlands, Erasmus University Medical Center, Rotterdam 3015CN, The Netherlands. ⁸Radboud University, Department of Molecular Biology, Faculty of Science, 6525GA Nijmegen, The Netherlands. ⁹European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK. ¹⁰Department of Cancer Genetics, Institute for Cancer Research, Oslo University Hospital, The Norwegian Radium Hospital, Oslo 0310, Norway. ¹¹K. G. Jebsen Centre for Breast Cancer Research, Institute for Clinical Medicine, University of Oslo, Oslo 0310, Norway. ¹²Department of Computer Science, University of Oslo, Oslo, Norway. ¹³Gachon Institute of Genome Medicine and Science, Gachon University Gil Medical Center, Incheon, South Korea. ¹⁴Translational Research Lab, Centre Léon Bérard, 28, rue Laënnec, 69373 Lyon Cedex 08, France. ¹⁵Department of Pathology, Brigham and Women's Hospital, Boston, Massachusetts 02115, USA. ¹⁶The Netherlands Cancer Institute, 1066 CX Amsterdam, The Netherlands. ¹⁷Breast Cancer Translational Research Laboratory, Université Libre de Bruxelles, Institut Jules Bordet, Bd de Waterloo 121, B-1000 Brussels, Belgium. ¹⁸Translational Cancer Research Unit, Center for Oncological Research, Faculty of Medicine and Health Sciences, University of Antwerp, Antwerp, Belgium. ¹⁹Dana-Farber Cancer Institute, Boston, Massachusetts 02215, USA. ²⁰Department of Pathology, Academic Medical Center, Meibergdreef 9, 1105 AZ Amsterdam, The Netherlands. ²¹Department of Pathology, Asan Medical Center, College of Medicine, Ulsan University, Ulsan, South Korea. ²²Department of Pathology, College of Medicine, Hanyang University, Seoul 133-791, South Korea. ²³Memorial Sloan Kettering Cancer Center, 1275 York Avenue, New York, New York 10065, USA. ²⁴Morgan Welch Inflammatory Breast Cancer Research Program and Clinic, The University of Texas MD Anderson Cancer Center, 1515 Holcombe Boulevard, Houston, Texas 77030, USA. ²⁵Institute for Bioengineering and Biopharmaceutical Research (IBBR), Hanyang University, Seoul, South Korea. ²⁶Institut National du Cancer, Research Division, Clinical Research Department, 52 avenue Moritz, 92513 Boulogne-Billancourt, France. ²⁷University Hospital of Minjoo, INSERM UMR 1098, Bd Fleming, Besançon 25000, France. ²⁸Pathology Department, Ninewells Hospital and Medical School, Dundee DD1 9SY, UK. ²⁹Oncologie Sénologie, ICM Institut Régional du Cancer, Montpellier, France. ³⁰The University of Queensland, UQ Centre for Clinical Research and School of Medicine, Brisbane, Queensland 4029, Australia. ³¹Cancer Research Laboratory, Faculty of Medicine, University of Iceland, 101 Reykjavik, Iceland. ³²IRCCS Istituto Tumori "Giovanni Paolo II", Bari, Italy. ³³Department of Pathology, Centre Léon Bérard, 28 rue Laënnec, 69373 Lyon Cedex 08, France. ³⁴Department of Pathology, GZA Hospitals Sint-Augustinus, Antwerp, Belgium. ³⁵Institut Curie, Paris Sciences Lettres University, Department of Pathology and INSERM U934, 26 rue d'Ulm, 75248 Paris Cedex 05, France. ³⁶Cancer Research UK Cambridge Institute, University of Cambridge, Li Ka Shing Centre, Robinson Way, Cambridge CB2 0RE, UK. ³⁷Breast Cancer Now Research Unit, King's College London, London SE1 9RT, UK. ³⁸Breast Cancer Now Toby Robins Research Centre, Institute of Cancer Research, London SW3 6JB, UK. ³⁹Department of Clinical Science, University of Bergen, 5020 Bergen, Norway. ⁴⁰Department of Oncology, Haukeland University Hospital, 5021 Bergen, Norway. ⁴¹National Cancer Centre Singapore, 11 Hospital Drive, 169610, Singapore. ⁴²Singapore General Hospital, Outram Road, 169608, Singapore. ⁴³Equipe Erable, INRIA Grenoble-Rhône-Alpes, 655, Avenue de l'Europe, 38330 Montbonnot-Saint Martin, France. ⁴⁴Synergie Lyon Cancer, Centre Léon Bérard, 28 rue Laënnec, Lyon Cedex 08, France. ⁴⁵Department of Genomic Medicine, UT MD Anderson Cancer Center, Houston, Texas 77230, USA. ⁴⁶Department of Radiation Oncology, Department of Laboratory Medicine, Radboud University Medical Center, Nijmegen 6525GA, The Netherlands. ⁴⁷Pathology Queensland, The Royal Brisbane and Women's Hospital, Brisbane, Queensland 4029, Australia. ⁴⁸Department of Breast Surgical Oncology, University of Texas MD Anderson Cancer Center, 1400 Pressler Street, Houston, Texas 77030, USA.

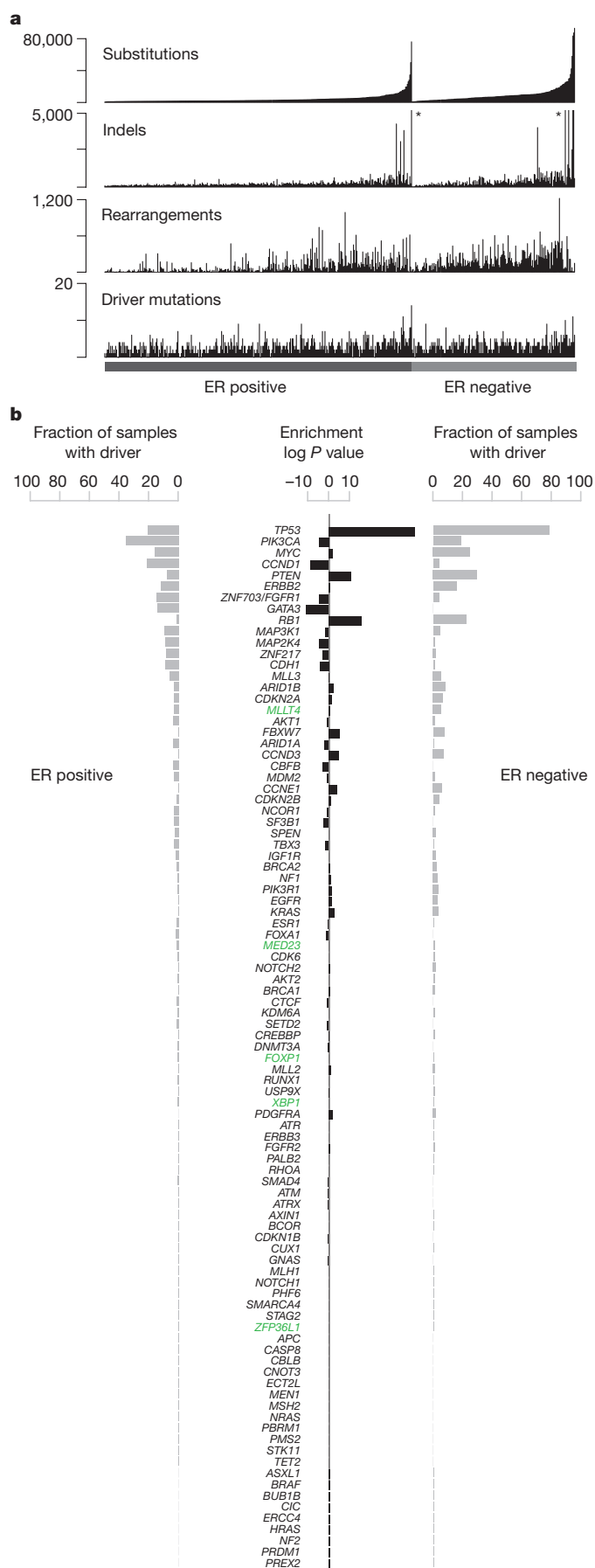


Figure 1 | Cohort and catalogue of somatic mutations in 560 breast cancers. **a**, Catalogue of base substitutions, insertions/deletions, rearrangements and driver mutations in 560 breast cancers (sorted by total substitution burden). Indel axis limited to 5,000(*). **b**, Complete list of curated driver genes sorted by frequency (descending). Fraction of ER-positive (left, total 366) and ER-negative (right, total 194) samples carrying a mutation in the relevant driver gene presented in grey. $\log_{10} P$ value of enrichment of each driver gene towards the ER-positive or ER-negative cohort is provided in black. Highlighted in green are genes for which there is new or further evidence supporting these as novel breast cancer genes.

and defective DNA maintenance, are responsible for generating these mutations^{1–3}.

Over the past five decades, several waves of technology have advanced the characterization of mutations in cancer genomes. Karyotype analysis revealed rearranged chromosomes and copy number alterations. Subsequently, loss of heterozygosity analysis, hybridization of cancer-derived DNA to microarrays and other approaches provided higher resolution insights into copy number changes^{4–8}. Recently, DNA sequencing has enabled systematic characterization of the full repertoire of mutation types including base substitutions, small insertions/deletions, rearrangements and copy number changes^{9–13}, yielding substantial insights into the mutated cancer genes and mutational processes operative in human cancer.

As for many cancer classes, most currently available breast cancer genome sequences target protein-coding exons^{8,11–15}. Consequently, there has been limited consideration of mutations in untranslated, intronic and intergenic regions, leaving central questions pertaining to the molecular pathogenesis of the disease unresolved. First, the role of activating driver rearrangements^{16–18} forming chimaeric (fusion) genes/proteins or relocating genes adjacent to new regulatory regions as observed in haematological and other malignancies¹⁹. Second, the role of driver substitutions and indels in non-coding regions of the genome^{20,21}. Common inherited variants conferring susceptibility to human disease are generally in non-coding regulatory regions and the possibility that similar mechanisms operate somatically in cancer was highlighted by the discovery of somatic driver substitutions in the *TERT* gene promoter^{22,23}. Third, which mutational processes generate the somatic mutations found in breast cancer^{2,24}. Addressing this question has been constrained because exome sequences do not inform on genome rearrangements and capture relatively few base substitution mutations, thus limiting statistical power to extract the mutational signatures imprinted on the genome by these processes^{24,25}.

Here we analyse whole-genome sequences of 560 cases in order to address these and other questions and to pave the way to a comprehensive understanding of the origins and consequences of somatic mutations in breast cancer.

Cancer genes and driver mutations

The whole genomes of 560 breast cancers and non-neoplastic tissue from each individual (556 female and 4 male) were sequenced (Supplementary Fig. 1, Supplementary Table 1). We detected 3,479,652 somatic base substitutions, 371,993 small indels and 77,695 rearrangements, with substantial variation in the number of each between individual samples (Fig. 1a, Supplementary Table 3). Transcriptome sequence, microRNA expression, array-based copy number and DNA methylation data were obtained from subsets of cases.

To identify new cancer genes, we combined somatic substitutions and indels in protein-coding exons with data from other series^{12–15,26}, constituting a total of 1,332 breast cancers, and searched for mutation clustering in each gene beyond that expected by chance. Five cancer genes were found for which evidence was previously absent or equivocal (*MED23*, *FOXP1*, *MLL4*, *XBPI*, *ZFP36L1*), or for which the mutations indicate the gene acts in breast cancer in a recessive rather than in a dominant fashion, as previously reported in other cancer types (see Supplementary Methods section 7.4 for detailed descriptions). From published reports on all cancer types (<http://cancer.sanger.ac.uk/census>),

we then compiled a list of 727 human cancer genes (Supplementary Table 12). On the basis of driver mutations found previously, we defined conservative rules for somatic driver base substitutions and indel mutations in each gene and sought mutations conforming to these rules in the 560 breast cancers. We identified 916 probable driver mutations of these classes (Fig. 1b, Supplementary Table 14, Extended Data Fig. 1).

To explore the role of genomic rearrangements as driver mutations^{16,18,19,27}, we sought predicted in-frame fusion genes that might create activated, dominant cancer genes. We identified 1,278 unique and 39 infrequently recurrent in-frame gene fusions (Supplementary Table 15). Many of the latter, however, were in regions of high rearrangement density, including amplicons²⁸ and fragile sites, and their recurrence is probably attributable to chance²⁷. Furthermore, transcriptome sequences from 260 cancers did not show expression of these fusions and generally confirmed the rarity of recurrent in-frame fusion genes. By contrast, recurrent rearrangements interrupting the gene footprints of *CDKN2A*, *RB1*, *MAP3K1*, *PTEN*, *MAP2K4*, *ARID1B*, *FBXW7*, *MLLT4* and *TP53* were found beyond the numbers expected from local background rearrangement rates, indicating that they contribute to the driver mutation burden of recessive cancer genes. Several other recurrently rearranged genomic regions were observed, including dominantly acting cancer genes *ETV6* and *ESR1* (without consistent elevation in expression levels), L1 retrotransposition sites²⁹ and fragile sites. The significance of these recurrently rearranged regions remains unclear (Extended Data Fig. 2).

Incorporation of recurrent copy number changes, including homozygous deletions and amplifications, generated a total of 1,628 likely driver mutations in 93 cancer genes (Fig. 1b). At least one driver was identifiable in 95% of cancers. The 10 most frequently mutated genes were *TP53*, *PIK3CA*, *MYC*, *CCND1*, *PTEN*, *ERBB2*, *ZNF703/FGFR1* locus, *GATA3*, *RB1* and *MAP3K1* (Fig. 1b, Extended Data Fig. 1), and these accounted for 62% of drivers.

Recurrent somatic mutations in non-coding regions

To investigate non-coding somatic driver substitutions and indels, we searched for non-coding genomic regions with more mutations than expected by chance (Fig. 2a, Supplementary Table 16, Extended Data Fig. 3).

The promoter of *PLEKHS1* exhibited recurrent mutations at two genomic positions³⁰ (Fig. 2a) TTTTGCAAT TGAACA ATTGCAAAA (as previously reported³⁰). The two mutated bases, within a 6 base pair (bp) core motif, are flanked, on either side by 9 base pairs of palindromic sequence forming inverted repeats³¹. Most cancers with these mutations showed many base substitutions of mutational signatures 2 and 13 that have been attributed to activity of APOBEC DNA-editing proteins that target the TCN sequence motif. One of the mutated bases is a cytosine in a TCA sequence context (shown above as the reverse complement, TGA) at which predominantly C>T substitutions were found. The other is a cytosine in ACA context, which showed both C>T and C>G mutations.

The TGAACA core sequence was mutated at the same two positions at multiple locations elsewhere in the genome (Supplementary Table 16c) where the TGAACA core was also flanked by palindromes albeit of different sequences and lengths (Supplementary Table 16c). These mutations were also usually found in cancers with many signature 2 and 13 mutations (Fig. 2a). TGAACA core sequences with longer flanking palindromes generally exhibited a higher mutation rate, and TGAACA sequences flanked by 9 bp palindromes exhibited an ~265-fold higher mutation rate than sequences without them (Fig. 2b, Supplementary Table 16d). However, additional factors must influence the mutation rate because it varied markedly between TGAACA core sequences with different palindromes of the same length (Fig. 2c). Some TGAACA-inverted repeat sites were in regulatory regions but others were intronic or intergenic without functional annotation (examples in Supplementary Table 16c) or exonic. The propensity for mutation recurrence at specific positions in a distinctive sequence motif in cancers with numerous mutations of particular signatures renders it plausible that these are hypermutable hotspots^{32–34}, perhaps through formation of DNA hairpin structures³⁵, which are single-stranded at their tips enabling attack by APOBEC enzymes, rather than driver mutations.

Two recurrently mutated sites were also observed in the promoter of *TBC1D12* (TBC1 domain family, member 12) (q value 4.5×10^{-2}) (Fig. 2a). The mutations were characteristic of signatures 2 and 13 and enriched in cancers with many signature 2 and 13 mutations (Fig. 2a). The mutations were within the *TBC1D12* Kozak consensus sequence

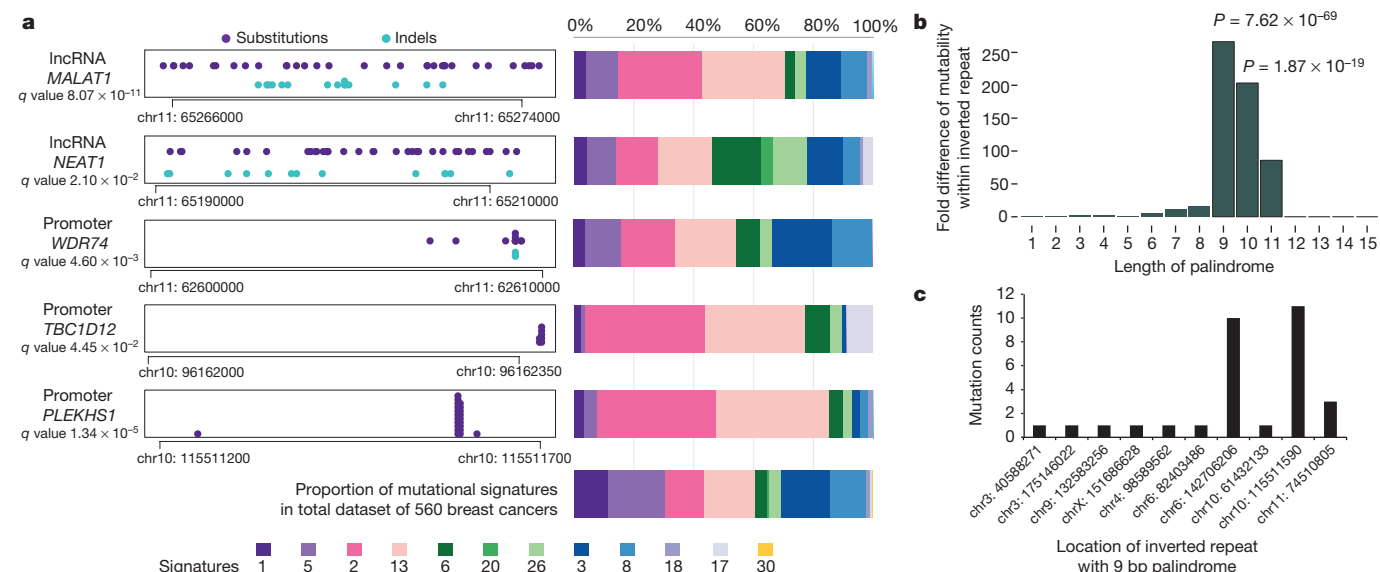


Figure 2 | Non-coding analyses of breast cancer genomes.

a, Distributions of substitution (purple dots) and indel (blue dots) mutations within the footprint of five regulatory regions identified as being more significantly mutated than expected is provided on the left. The proportion of base substitution mutation signatures associated with corresponding samples carrying mutations in each of these non-coding

regions, is displayed on the right. **b**, Mutability of TGAACA/TGTTCA motifs within inverted repeats of varying flanking palindromic sequence length compared to motifs not within an inverted repeat. **c**, Variation in mutability between loci of TGAACA/TGTTCA inverted repeats with 9 bp palindromes.

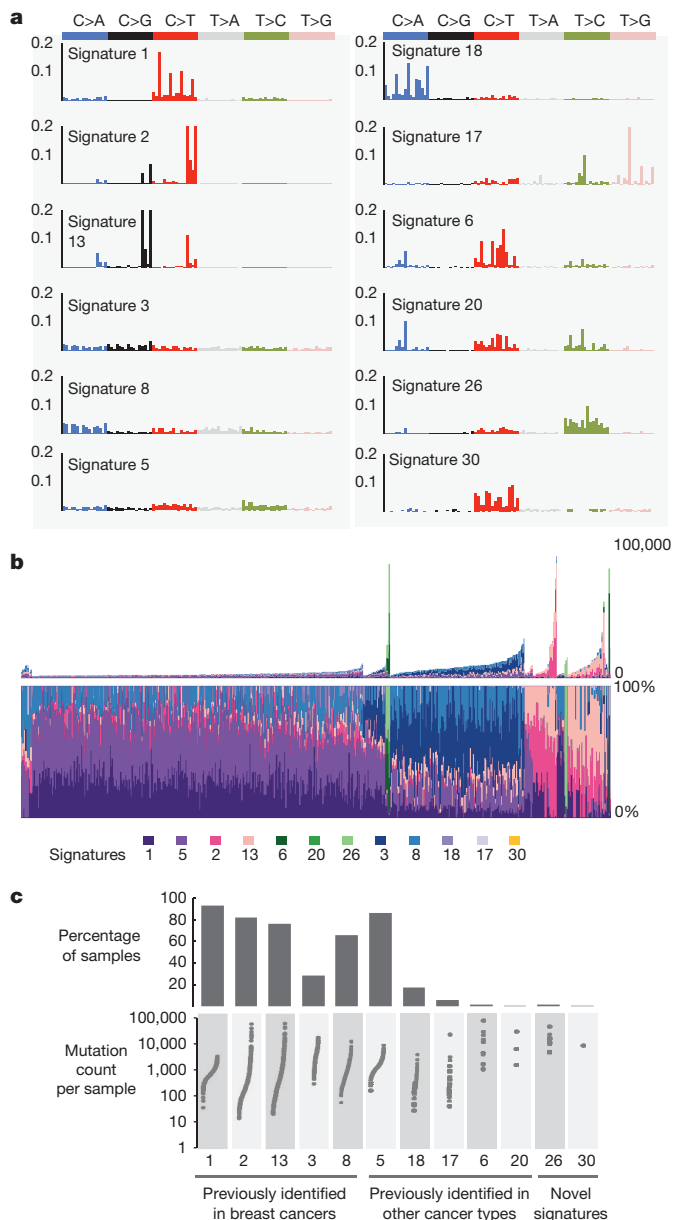


Figure 3 | Extraction and contributions of base substitution signatures in 560 breast cancers. **a**, Twelve mutation signatures extracted using non-negative matrix factorization. Each signature is ordered by mutation class (C>A/G>T, C>G/G>C, C>T/G>A, T>A/A>T, T>C/A>G, T>G/A>C), taking immediate flanking sequence into account. For each class, mutations are ordered by 5' base (A, C, G, T) first before 3' base (A, C, G, T). **b**, The spectrum of base substitution signatures within 560 breast cancers. Mutation signatures are ordered (and coloured) according to broad biological groups: signatures 1 and 5 are correlated with age of diagnosis; signatures 2 and 13 are putatively APOBEC-related; signatures 6, 20 and 26 are associated with mismatch-repair deficiency; signatures 3 and 8 are associated with homologous-recombination deficiency; signatures 18, 17 and 30 have unknown aetiology. For ease of reading, this arrangement is adopted for the rest of the manuscript. Samples are ordered according to hierarchical clustering performed on mutation signatures. Top, absolute numbers of mutations of each signature in each sample. Bottom, proportion of each signature in each sample. **c**, Distribution of mutation counts for each signature in relevant breast cancer samples. Percentage of samples carrying each signature provided above each signature.

(CCCCAGATGGTGGG)), shifting it away from the consensus³⁶. The association with particular mutational signatures suggests that these may also be in a region of hypermutability rather than drivers.

The *WDR74* promoter showed base substitutions and indels (q value 4.6×10^{-3}) forming a cluster of overlapping mutations²⁰ (Fig. 2a). Coding sequence driver mutations in *WDR74* have not been reported. No differences were observed in *WDR74* transcript levels between cancers with *WDR74* promoter mutations compared to those without. Nevertheless, the pattern of this non-coding mutation cluster, with overlapping and different mutation types, is more compatible with the possibility of the mutations being drivers.

Two long non-coding RNAs, *MALAT1* (q value 8.7×10^{-11} , as previously reported¹²) and *NEAT1* (q value 2.1×10^{-2}) were enriched with mutations. Transcript levels were not significantly different between mutated and non-mutated samples. Whether these mutations are drivers or result from local hypermutability is unclear.

Mutational signatures

Mutational processes generating somatic mutations imprint particular patterns of mutations on cancer genomes, termed signatures^{2,24,37}. Applying a mathematical approach²⁵ to extract mutational signatures previously revealed five base-substitution signatures in breast cancer: signatures 1, 2, 3, 8 and 13 (refs 2, 24). Using this method for the 560 cases revealed twelve signatures, including those previously observed and a further seven, of which five have formerly been detected in other cancer types (signatures 5, 6, 17, 18 and 20) and two are new (signatures 26 and 30) (Fig. 3a, b, 4a, Supplementary Table 21a–c, Supplementary Methods section 15). Two indel signatures were also found^{2,24}.

Signatures of rearrangement mutational processes have not previously been formally investigated. To enable this we adopted a rearrangement classification incorporating 32 subclasses. In many cancer genomes, large numbers of rearrangements are regionally clustered, for example in zones of gene amplification. Therefore, we first classified rearrangements into those inside and outside clusters, further subclassified them into deletions, inversions and tandem duplications, and then according to the size of the rearranged segment. The final category in both groups was interchromosomal translocations.

Application of the mathematical framework used for base substitution signatures^{2,24,25} extracted six rearrangement signatures (Fig. 4b, Supplementary Table 21). Unsupervised hierarchical clustering on the basis of the proportion of rearrangements attributed to each signature in each breast cancer yielded seven major subgroups exhibiting distinct associations with other genomic, histological or gene expression features (Fig. 5, Extended Data Figs 4–6).

Rearrangement signature 1 (9% of all rearrangements) and rearrangement signature 3 (18% rearrangements) were characterized predominantly by tandem duplications (Fig. 4b). Tandem duplications associated with rearrangement signature 1 were mostly >100 kb (Fig. 4b), and those with rearrangement signature 3 were <10 kb (Fig. 4b, Extended Data Fig. 7). More than 95% of rearrangement signature 3 tandem duplications were concentrated in 15% of cancers (cluster D, Fig. 5), many with several hundred rearrangements of this type. Almost all cancers (91%) with *BRCA1* mutations or promoter hypermethylation were in this group, which was enriched for basal-like, triple negative cancers and copy number classification of a high homologous recombination deficiency (HRD) index^{38–40}. Thus, inactivation of *BRCA1*, but not *BRCA2*, may be responsible for the rearrangement signature 3 small tandem duplication mutator phenotype.

More than 35% of rearrangement signature 1 tandem duplications were found in just 8.5% of the breast cancers and some cases had hundreds of these (cluster F, Fig. 5). The cause of this large tandem duplication mutator phenotype (Fig. 4b) is unknown. Cancers exhibiting it are frequently TP53-mutated, relatively late diagnosis, triple-negative breast cancers, showing enrichment for base substitution signature 3 and a high HRD index (Fig. 5), but do not have *BRCA1/2* mutations or *BRCA1* promoter hypermethylation.

Rearrangement signature 1 and 3 tandem duplications (Extended Data Fig. 7) were generally evenly distributed over the genome. However,

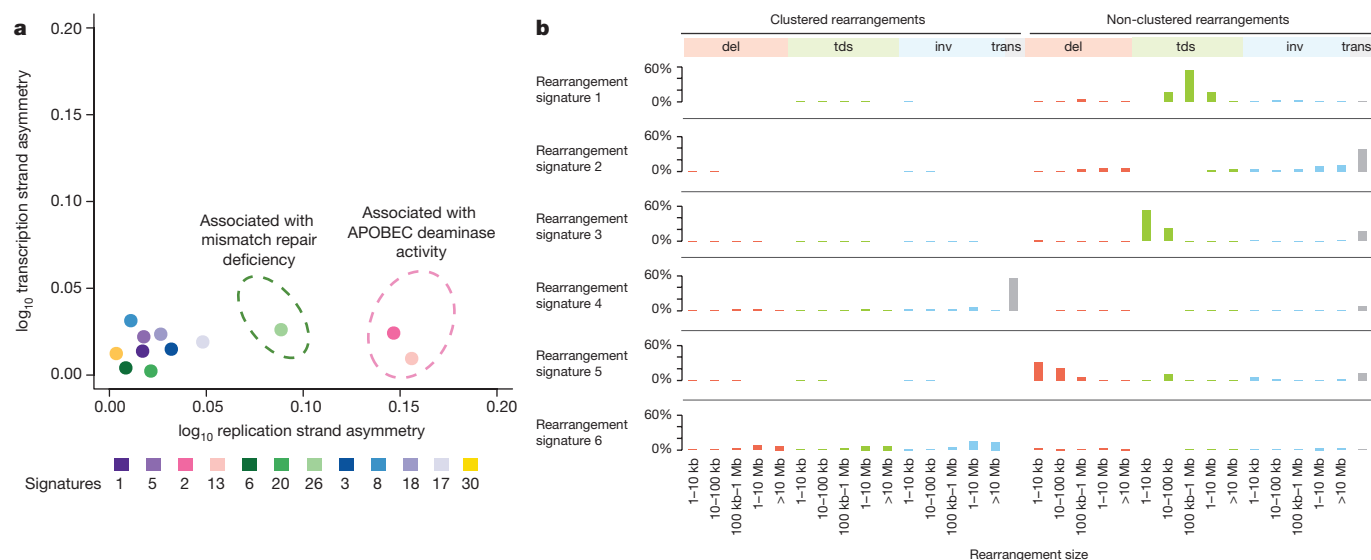


Figure 4 | Additional characteristics of base substitution signatures and novel rearrangement signatures in 560 breast cancers. a, Contrasting transcriptional strand asymmetry and replication strand asymmetry between twelve base substitution signatures. **b,** Six rearrangement

there were nine locations at which recurrence of tandem duplications was found across the breast cancers and which often showed multiple, nested tandem duplications in individual cases (Extended Data Fig. 8). These may be mutational hotspots specific for these tandem duplication mutational processes, although we cannot exclude the possibility that they represent driver events.

Rearrangement signature 5 (accounting for 14% rearrangements) was characterized by deletions <100 kb. It was strongly associated with the presence of *BRCA1* mutations or promoter hypermethylation (cluster D, Fig. 5), *BRCA2* mutations (cluster G, Fig. 5) and with rearrangement signature 1 large tandem duplications (cluster F, Fig. 5).

Rearrangement signature 2 (accounting for 22% rearrangements) was characterized by non-clustered deletions (>100 kb), inversions and interchromosomal translocations, was present in most cancers but was particularly enriched in oestrogen receptor (ER)-positive cancers with quiet copy number profiles (cluster E, GISTIC (genomic identification of significant targets in cancer) cluster 3; Fig. 5). Rearrangement signature 4 (accounting for 18% of rearrangements) was characterized by clustered interchromosomal translocations, whereas rearrangement signature 6 (19% of rearrangements) had clustered inversions and deletions (clusters A, B, C; Fig. 5).

Short segments (1–5 bp) of overlapping microhomology characteristic of alternative methods of end-joining repair were found at most rearrangements^{2,14}. Rearrangement signatures 2, 4 and 6 were characterized by a peak at 1 bp of microhomology, whereas rearrangement signatures 1, 3 and 5, associated with homologous recombination DNA repair deficiency, exhibited a peak at 2 bp (Extended Data Fig. 9). Thus, different end-joining mechanisms may operate with different rearrangement processes. A proportion of breast cancers showed rearrangement signature 5 deletions with longer (>10 bp) microhomologies involving sequences from short-interspersed nuclear elements, most commonly AluS (63%) and AluY (15%) family repeats (Extended Data Fig. 9). Long segments (more than 10 bp) of non-templated sequence were particularly enriched amongst clustered rearrangements.

Localized hypermutation: kataegis

Focal base-substitution hypermutation, termed kataegis, is generally characterized by substitutions with characteristic features of signatures 2 and 13 (refs 2, 24). Kataegis was observed in 49% breast cancers, with 4% exhibiting 10 or more foci (Supplementary Table 21c). Kataegis colocalizes with clustered rearrangements characteristic of rearrangement

signatures extracted using non-negative matrix factorization. Probability of rearrangement element on y axis. Rearrangement size on x axis. del, deletion; tds, tandem duplication; inv, inversion; trans, translocation.

signatures 4 and 6 (Fig. 4b). Cancers with tandem duplications or deletions of rearrangement signatures 1, 3 and 5 did not usually demonstrate kataegis. However, there must be additional determinants of kataegis as only 2% of rearrangements are associated with it. A rare (14 out of 1,557 foci, 0.9%) alternative form of kataegis, colocalizing with rearrangements but with a base-substitution pattern characterized by T>G and T>C mutations, predominantly at NTT and NTA sequences (where N can be any base A, T, C or G), was also observed (Extended Data Fig. 10). This pattern of base substitutions most closely matches signature 9 (Extended Data Fig. 10; <http://cancer.sanger.ac.uk/cosmic/signatures>), previously observed in B lymphocyte neoplasms and attributed to polymerase eta activity⁴¹.

Mutational signatures exhibit distinct DNA replication strand biases

The distributions of mutations attributable to each of the 20 mutational signatures (12 base substitution, 2 indel and 6 rearrangement) were explored⁴² with respect to DNA replication strand. We found an asymmetric distribution of mutations between leading and lagging replication strands for many, but not all signatures⁴² (Fig. 4a). Notably, signatures 2 and 13, owing to APOBEC deamination, showed marked lagging-strand replication bias (Fig. 4a) suggesting that lagging-strand replication provides single-stranded DNA for APOBEC deamination. Of the three signatures associated with mismatch-repair deficiency (signatures 6, 20 and 26), only signature 26 exhibited replicative-strand bias, highlighting how different signatures arising from defects of the same pathway can exhibit distinct relationships with replication.

Mutational signatures associated with *BRCA1* and *BRCA2* mutations

Of the 560 breast cancers, 90 had germline (60) or somatic (14) inactivating mutations in *BRCA1* (35) or *BRCA2* (39) or showed methylation of the *BRCA1* promoter (16). Loss of the wild-type chromosome 17 or 13 was observed in 80 out of 90 cases. The latter exhibited many base substitution mutations of signature 3, accompanied by deletions of >3 bp with microhomology at rearrangement break-points, and signature 8 together with CC>AA double nucleotide substitutions. Cases in which the wild-type chromosome 17 or 13 was retained did not show these signatures. Thus signature 3 and, to a lesser extent, signature 8 are associated with absence of *BRCA1* and *BRCA2* functions.

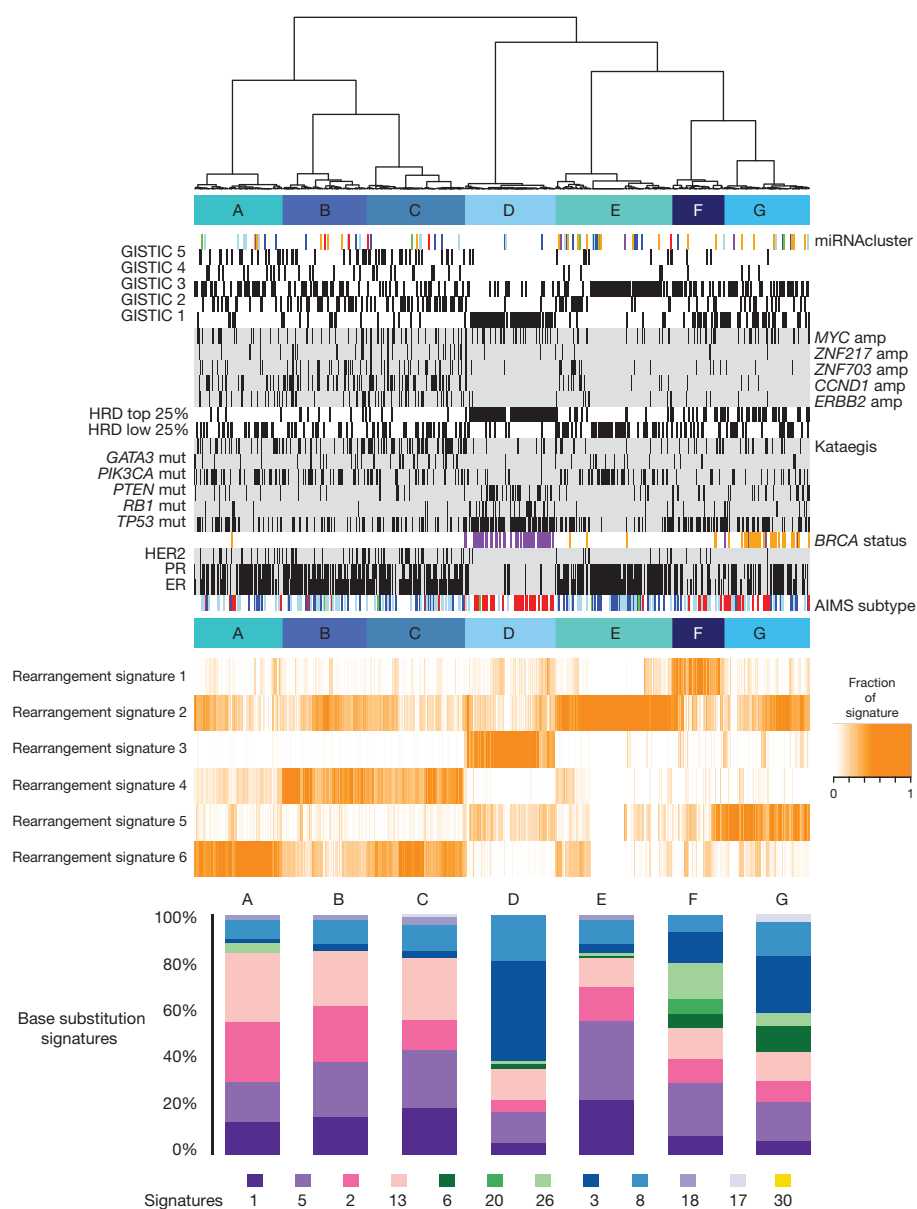


Figure 5 | Integrative analysis of rearrangement signatures. Heatmap of rearrangement signatures following unsupervised hierarchical clustering based on proportions of rearrangement signatures in each cancer. Seven cluster groups (A–G) noted and relationships with expression (AIMS) subtype (basal, red; luminal B, light blue; luminal A, dark blue), immunohistopathology status (ER, progesterone receptor (PR), HER2 status; black, positive), abrogation of *BRCA1* (purple) and *BRCA2* (orange) (whether germline, somatic or through promoter hypermethylation), presence of 3 or more foci of kataegis (black, positive), HRD index (top 25% or lowest 25%; black, positive), GISTIC cluster group (black, positive) and driver mutations in cancer genes. miRNA cluster groups: 0, red; 1, purple; 2, blue; 3, light blue; 4, green; 5, orange. Contribution of base-substitution signatures in these seven cluster groups is provided in the bottom panel.

Cancers with inactivating *BRCA1* or *BRCA2* mutations usually carry many genomic rearrangements. Cancers with *BRCA1*, but not *BRCA2*, mutations exhibit large numbers of rearrangement signature 3 small tandem duplications. Cancers with *BRCA1* or *BRCA2* mutations show substantial numbers of rearrangement signature 5 deletions. No other rearrangement signatures were associated with *BRCA1*- or *BRCA2*-null cases (clusters D and G, Fig. 5). Some breast cancers without identifiable *BRCA1/2* mutations or *BRCA1* promoter methylation showed these features and segregated with *BRCA1/2*-null cancers in hierarchical clustering analysis (Fig. 5). In such cases, the *BRCA1/2* mutations may have been missed or other mutated or promoter methylated genes may be exerting similar effects (see <http://cancer.sanger.ac.uk/cosmic/sample/genomes> for examples of whole-genome profiles of typical *BRCA1*-null, (for example, PD6413a, PD7215a) and *BRCA2*-null tumours (for example, PD4952a, PD4955a)).

A further subset of cancers (cluster F, Fig. 5) show similarities in mutational pattern to *BRCA1/2*-null cancers, with many rearrangement signature 5 deletions and enrichment for base substitution signatures 3 and 8. However, these do not segregate together with *BRCA1/2*-null cases in hierarchical clustering analysis, have rearrangement signature 1 large tandem duplications and do not show *BRCA1/2* mutations. Somatic and germline mutations in genes associated with the DNA

double-strand break repair pathway including *ATM*, *ATR*, *PALB2*, *RAD51C*, *RAD50*, *TP53*, *CHEK2* and *BRIP1*, were sought in these cancers. We did not observe any clear-cut relationships between mutations in these genes and these mutational patterns.

Cancers with *BRCA1/2* mutations are particularly responsive to cisplatin and PARP inhibitors^{43–45}. Combinations of base substitution, indel and rearrangement mutational signatures may be better biomarkers of defective homologous-recombination-based DNA double-strand break repair and responsiveness to these drugs⁴⁶ than *BRCA1/2* mutations or promoter methylation alone and thus may constitute the basis of future diagnostics.

Conclusions

A comprehensive perspective on the somatic genetics of breast cancer is drawing closer (see <http://cancer.sanger.ac.uk/cosmic/sample/genomes> for individual patient genome profile, and Methods for orientation). At least 12 base substitution mutational signatures and 6 rearrangement signatures contribute to the somatic mutations found, and 93 mutated cancer genes (31 dominant, 60 recessive, 2 uncertain) are implicated in genesis of the disease. However, dominantly acting activated fusion genes and non-coding driver mutations appear rare. Additional infrequently mutated cancer genes probably exist.

However, the genes harbouring the substantial majority of driver mutations are now known.

Nevertheless, important questions remain to be addressed. Recurrent mutational events including whole-chromosome copy number changes and unexplained regions with recurrent rearrangements could harbour additional cancer genes. Identifying non-coding drivers is challenging and requires further investigation. Although almost all breast cancers have at least one identifiable driver mutation, the number with only a single identified driver is perhaps surprising. The roles of viruses or other microbes have not been exhaustively examined. Thus, further exploration and analysis of whole-genome sequences from breast cancer patients will be required to complete our understanding of the somatic mutational basis of the disease.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 29 June 2015; accepted 17 March 2016.

Published online 2 May 2016.

1. Stratton, M. R., Campbell, P. J. & Futreal, P. A. The cancer genome. *Nature* **458**, 719–724 (2009).
2. Nik-Zainal, S. *et al.* Mutational processes molding the genomes of 21 breast cancers. *Cell* **149**, 979–993 (2012).
3. Nik-Zainal, S. *et al.* The life history of 21 breast cancers. *Cell* **149**, 994–1007 (2012).
4. Hicks, J. *et al.* Novel patterns of genome rearrangement and their association with survival in breast cancer. *Genome Res.* **16**, 1465–1479 (2006).
5. Bergamaschi, A. *et al.* Extracellular matrix signature identifies breast cancer subgroups with different clinical outcome. *J. Pathol.* **214**, 357–367 (2008).
6. Ching, H. C., Naidu, R., Seong, M. K., Har, Y. C. & Taib, N. A. Integrated analysis of copy number and loss of heterozygosity in primary breast carcinomas using high-density SNP array. *Int. J. Oncol.* **39**, 621–633 (2011).
7. Fang, M. *et al.* Genomic differences between estrogen receptor (ER)-positive and ER-negative human breast carcinoma identified by single nucleotide polymorphism array comparative genome hybridization analysis. *Cancer* **117**, 2024–2034 (2011).
8. Curtis, C. *et al.* The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature* **486**, 346–352 (2012).
9. Pleasance, E. D. *et al.* A comprehensive catalogue of somatic mutations from a human cancer genome. *Nature* **463**, 191–196 (2010).
10. Pleasance, E. D. *et al.* A small-cell lung cancer genome with complex signatures of tobacco exposure. *Nature* **463**, 184–190 (2010).
11. Banerji, S. *et al.* Sequence analysis of mutations and translocations across breast cancer subtypes. *Nature* **486**, 405–409 (2012).
12. Ellis, M. J. *et al.* Whole-genome analysis informs breast cancer response to aromatase inhibition. *Nature* **486**, 353–360 (2012).
13. Shah, S. P. *et al.* The clonal and mutational evolution spectrum of primary triple-negative breast cancers. *Nature* **486**, 395–399 (2012).
14. Stephens, P. J. *et al.* The landscape of cancer genes and mutational processes in breast cancer. *Nature* **486**, 400–404 (2012).
15. The Cancer Genome Atlas Network Comprehensive molecular portraits of human breast tumours. *Nature* **490**, 61–70 (2012).
16. Wu, Y. M. *et al.* Identification of targetable FGFR gene fusions in diverse cancers. *Cancer Discovery* **3**, 636–647 (2013).
17. Giacomini, C. P. *et al.* Breakpoint analysis of transcriptional and genomic profiles uncovers novel gene fusions spanning multiple human cancer types. *PLoS Genet.* **9**, e1003464 (2013).
18. Robinson, D. R. *et al.* Functionally recurrent rearrangements of the MAST kinase and Notch gene families in breast cancer. *Nature Med.* **17**, 1646–1651 (2011).
19. Karlsson, J. *et al.* Activation of human telomerase reverse transcriptase through gene fusion in clear cell sarcoma of the kidney. *Cancer Lett.* **357**, 498–501 (2015).
20. Khurana, E. *et al.* Integrative annotation of variants from 1092 humans: application to cancer genomics. *Science* **342**, 1235587 (2013).
21. West, J. A. *et al.* The long noncoding RNAs NEAT1 and MALAT1 bind active chromatin sites. *Mol. Cell* **55**, 791–802 (2014).
22. Huang, F. W. *et al.* Highly recurrent *TERT* promoter mutations in human melanoma. *Science* **339**, 957–959 (2013).
23. Vinagre, J. *et al.* Frequency of *TERT* promoter mutations in human cancers. *Nature Commun.* **4**, 2185 (2013).
24. Alexandrov, L. B. *et al.* Signatures of mutational processes in human cancer. *Nature* **500**, 415–421 (2013).
25. Alexandrov, L. B., Nik-Zainal, S., Wedge, D. C., Campbell, P. J. & Stratton, M. R. Deciphering signatures of mutational processes operative in human cancer. *Cell Rep.* **3**, 246–259 (2013).
26. Lawrence, M. S. *et al.* Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature* **505**, 495–501 (2014).
27. Natrajan, R. *et al.* Characterization of the genomic features and expressed fusion genes in micropapillary carcinomas of the breast. *J. Pathol.* **232**, 553–565 (2014).
28. Kalyana-Sundaram, S. *et al.* Gene fusions associated with recurrent amplicons represent a class of passenger aberrations in breast cancer. *Neoplasia* **14**, 702–708 (2012).
29. Tubio, J. M. Somatic structural variation and cancer. *Brief. Funct. Genomics* **14**, 339–351 (2015).
30. Weinhold, N., Jacobsen, A., Schultz, N., Sander, C. & Lee, W. Genome-wide analysis of noncoding regulatory mutations in cancer. *Nature Genet.* **46**, 1160–1165 (2014).
31. Ussery, D. W., Binnewies, T. T., Gouveia-Oliveira, R., Jarmer, H. & Hallin, P. F. Genome update: DNA repeats in bacterial genomes. *Microbiology* **150**, 3519–3521 (2004).
32. Lu, S. *et al.* Short inverted repeats are hotspots for genetic instability: relevance to cancer genomes. *Cell Rep.* **10**, 1674–1680 (2015).
33. Voineagu, I., Narayanan, V., Lobachev, K. S. & Mirkin, S. M. Replication stalling at unstable inverted repeats: interplay between DNA hairpins and fork stabilizing proteins. *Proc. Natl Acad. Sci. USA* **105**, 9936–9941 (2008).
34. Wojcik, E. A. *et al.* Direct and inverted repeats elicit genetic instability by both exploiting and eluding DNA double-strand break repair systems in mycobacteria. *PLoS ONE* **7**, e51064 (2012).
35. Pearson, C. E., Zorbas, H., Price, G. B. & Zannis-Hadjopoulos, M. Inverted repeats, stem-loops, and cruciforms: significance for initiation of DNA replication. *J. Cell. Biochem.* **63**, 1–22 (1996).
36. Kozak, M. Interpreting cDNA sequences: some insights from studies on translation. *Mamm. Genome* **7**, 563–574 (1996).
37. Helleday, T., Eshtad, S. & Nik-Zainal, S. Mechanisms underlying mutational signatures in human cancers. *Nature Rev. Genet.* **15**, 585–598 (2014).
38. Birkbak, N. J. *et al.* Telomeric allelic imbalance indicates defective DNA repair and sensitivity to DNA-damaging agents. *Cancer Disc.* **2**, 366–375 (2012).
39. Abkevich, V. *et al.* Patterns of genomic loss of heterozygosity predict homologous recombination repair defects in epithelial ovarian cancer. *Br. J. Cancer* **107**, 1776–1782 (2012).
40. Popova, T. *et al.* Ploidy and large-scale genomic instability consistently identify basal-like breast carcinomas with *BRCA1/2* inactivation. *Cancer Res.* **72**, 5454–5462 (2012).
41. Puente, X. S. *et al.* Whole-genome sequencing identifies recurrent mutations in chronic lymphocytic leukaemia. *Nature* **475**, 101–105 (2011).
42. Morganello, S. A. *et al.* The topography of mutational processes in breast cancer genomes. *Nature Commun.* <http://dx.doi.org/10.1038/ncomms11383> (2016).
43. Fong, P. C. *et al.* Inhibition of poly(ADP-ribose) polymerase in tumors from BRCA mutation carriers. *N. Engl. J. Med.* **361**, 123–134 (2009).
44. Forster, M. D. *et al.* Treatment with olaparib in a patient with PTEN-deficient endometrioid endometrial cancer. *Nature Rev. Clin. Oncol.* **8**, 302–306 (2011).
45. Turner, N., Tutt, A. & Ashworth, A. Targeting the DNA repair defect of BRCA tumours. *Curr. Opin. Pharmacol.* **5**, 388–393 (2005).
46. Waddell, N. *et al.* Whole genomes redefine the mutational landscape of pancreatic cancer. *Nature* **518**, 495–501 (2015).

Supplementary Information is available in the online version of the paper.

Acknowledgements This work has been funded through the ICGC Breast Cancer Working group by the Breast Cancer Somatic Genetics Study (BASIS), a European research project funded by the European Community's Seventh Framework Programme (FP7/2010-2014) under the grant agreement number 242006; the Triple Negative project funded by the Wellcome Trust (grant reference 077012/Z/05/Z) and the HER2+ project funded by Institut National du Cancer (INCa) in France (grant numbers 226-2009, 02-2011, 41-2012, 144-2008, 06-2012). The ICGC Asian Breast Cancer Project was funded through a grant of the Korean Health Technology R&D Project, Ministry of Health and Welfare, Republic of Korea (A11218-SC01). Personally funded by grants above: F.G.R.-G., S.M., K.R., S.M. were funded by BASIS. Recruitment was performed under the auspices of the ICGC breast cancer projects run by the UK, France and Korea. For contributions towards instruments, specimens and collections: Tayside Tissue Bank (funded by CRUK, University of Dundee, Chief Scientist Office & Breast Cancer Campaign), Asan Bio-Resource Center of the Korea Biobank Network, Seoul, South Korea, OSBREAC consortium, The Icelandic Centre for Research (RANNIS), The Swedish Cancer Society and the Swedish Research Council, and Fondation Jean Dausset-Centre d'Etudes du polymorphisme humain. Icelandic Cancer Registry, The Brisbane Breast Bank (The University of Queensland, The Royal Brisbane and Women's Hospital and QIMR Berghofer), Breast Cancer Tissue and Data Bank at KCL and NIHR Biomedical Research Centre at Guy's and St Thomas's Hospitals. Breakthrough Breast Cancer and Cancer Research UK Experimental Cancer Medicine Centre at KCL. For pathology review: The Mouse Genome Project and Department of Pathology, Cambridge University Hospitals NHS Foundation Trust for microscopes. A. Richardson, A. Ehinger, A. Vincent-Salomon, C. Van Deuren, C. Purdie, D. Larsimont, D. Giri, D. Grabau, E. Provenzano, G. MacGrogan, G. Van den Eynden, I. Treilleux, J. E. Brock, J. Jacquemier, J. Reis-Filho, L. Arnould, L. Jones, M. van de Vijver, Ø. Garred, R. Salgado, S. Pinder, S. R. Lakhani, T. Sauer, V. Barbashina. Illumina UK Ltd for input on optimization of sequencing throughout this project. Wellcome Trust Sanger Institute Sequencing Core Facility, Core IT Facility and Cancer Genome Project Core IT

team and Cancer Genome Project Core Laboratory team for general support. Personal funding: S.N.-Z. is a Wellcome Beit Fellow and personally funded by a Wellcome Trust Intermediate Fellowship (WT100183MA). L.B.A. is supported through a J. Robert Oppenheimer Fellowship at Los Alamos National Laboratory. A.L.R. is partially supported by the Dana-Farber/Harvard Cancer Center SPORC in Breast Cancer (NIH/NCI 5 P50 CA168504-02). D.G. was supported by the EU-FP7-SUPPRESSTEM project. A.S. was supported by Cancer Genomics Netherlands through a grant from the Netherlands Organisation of Scientific research (NWO). M.S. was supported by the EU-FP7-DDR response project. C.S. and C.D. are supported by a grant from the Breast Cancer Research Foundation. E.B. was funded by EMBL. C.S. is funded by FNRS (Fonds National de la Recherche Scientifique). S.J.J. is supported by Leading Foreign Research Institute Recruitment Program through the National Research Foundation of Republic Korea (NRF 2011-0030105). G.K. is supported by National Research Foundation of Korea (NRF) grants funded by the Korean government (NRF 2015R1A2A1A10052578). J.F. received funding from an ERC Advanced grant (no. 322737). For general contribution and administrative support: Fondation Synergie Lyon Cancer in France. J. G. Jonasson, Department of Pathology, University Hospital & Faculty of Medicine, University of Iceland. K. Ferguson, Tissue Bank Manager, Brisbane Breast Bank and The Breast Unit, The Royal Brisbane and Women's Hospital, Brisbane, Australia. The Oslo Breast Cancer Consortium of Norway (OSBREAC). Angelo Paradiso, IRCCS Istituto Tumori "Giovanni Paolo II", Bari Italy. A. Vines for administratively supporting to identifying the samples, organizing the bank, and sending out the samples. M. Schlooz-Vries, J. Tol, H. van Laarhoven, F. Sweep, P. Bult in Nijmegen for contributions in Nijmegen. This research used resources provided by the Los Alamos National Laboratory Institutional Computing Program, which is supported by the US Department of Energy National Nuclear Security Administration under contract no. DE-AC52-06NA25396. Research performed at Los Alamos National Laboratory was carried out under the auspices of the National Nuclear Security Administration of the United States Department of

Energy. N. Miller (in memoriam) for her contribution in setting up the clinical database. Finally, we would like to acknowledge all members of the ICGC Breast Cancer Working Group and ICGC Asian Breast Cancer Project.

Author Contributions S.N.-Z., M.R.S. designed the study, analysed data and wrote the manuscript. H.D., J.S., M. Ramakrishna, D.G., X.Z. performed curation of data and contributed towards genomic and copy number analyses. M.S., A.B.B., M.R.A., O.C.L., A.L., M. Ringner, contributed towards curation and analysis of non-genomic data (transcriptomic, miRNA, methylation). I.M., L.B.A., D.C.W., P.V.L., S. Morgarella, Y.S.J., contributed towards specialist analyses. G.T., G.K., A.L.R., A-L.B.-D., J.W.M.M., M.J.v.d.V., H.G.S., E.B., A. Borg., A.V., P.A.F., P.J.C., designed the study, drove the consortium and provided samples. S.Martin was the project coordinator. S.McL., S.O.M., K.R., contributed operationally. S.-M.A., S.B., J.E.B., A.Brooks., C.D., L.D., A.F., J.A.F., G.K.J.H., S.J.J., H.-Y.K., T.A.K., S.K., H.J.L., J.-Y.L., I.P., X.P., C.A.P., F.G.R.-G., G.R., A.M.S., P.T.S., O.A.S., S.T., I.T., G.G.V.d.E., P.V., A.V.-S., L.Y., C.C., L.v.V., A.T., S.K., B.K.T.T., J.J., N.t.U., C.S., P.N.S., S.V.L., S.R.L., J.E.E., A.M.T. contributed pathology assessment and/or samples. A. Butler, S.D., M.G., D.R.J., Y.L., A.M., V.M., K.R., R.S., L.S., J.T. contributed IT processing and management expertise. All authors discussed the results and commented on the manuscript.

Author Information Raw data have been submitted to the European-Genome Phenome Archive under the overarching accession number EGAS00001001178 (please see Supplementary Notes for breakdown by data type). Somatic variants have been deposited at the International Cancer Genome Consortium Data Portal (<https://dcc.icgc.org/>). Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to G.K. (gkong@hanyang.ac.kr), S.N.-Z. (snz@sanger.ac.uk), M.S. (mrs@sanger.ac.uk) or A.V. (Alain.Viari@inria.fr).

METHODS

Data reporting. No statistical methods were used to predetermine sample size. The experiments were not randomized and the investigators were not blinded to allocation during experiments and outcome assessment.

Sample selection. DNA was extracted from 560 breast cancers and normal tissue (peripheral blood lymphocytes, adjacent normal breast tissue or skin) and total RNA extracted from 268 of the same individuals. Samples were subjected to pathology review and only samples assessed as being composed of >70% tumour cells, were accepted for inclusion in the study (Supplementary Table 1).

Massively parallel sequencing and alignment. Short insert 500 bp genomic libraries and 350 bp poly-A-selected transcriptomic libraries were constructed, flowcells prepared and sequencing clusters generated according to Illumina library protocols⁴⁷. We performed 108 base/100 base (genomic), or 75 base (transcriptomic) paired-end sequencing on Illumina GAIIx, HiSeq 2000 or HiSeq 2500 genome analysers, in accordance with the Illumina Genome Analyzer operating manual. The average sequence coverage was 40.4-fold for tumour samples and 30.2-fold for normal samples (Supplementary Table 2).

Short insert paired-end reads were aligned to the reference human genome (GRCh37) using Burrows-Wheeler Aligner, BWA (v0.5.9)⁴⁸. RNA sequencing data was aligned to the human reference genome (GRCh37) using TopHat (v1.3.3) (<http://ccb.jhu.edu/software/tophat/index.shtml>).

Processing of genomic data. CaVEMan (Cancer Variants Through Expectation Maximization: <http://cancerit.github.io/CaVEMan/>) was used for calling somatic substitutions.

Indels in the tumour and normal genomes were called using a modified Pindel version 2.0. (<http://cancerit.github.io/cgpPindel/>) on the NCBI37 genome build⁴⁹.

Structural variants were discovered using a bespoke algorithm, BRASS (BReakpoint Analysis) (<https://github.com/cancerit/BRASS>) through discordantly mapping paired-end reads. Next, discordantly mapping read pairs that were likely to span breakpoints, as well as a selection of nearby properly paired reads, were grouped for each region of interest. Using the Velvet *de novo* assembler⁵⁰, reads were locally assembled within each of these regions to produce a contiguous consensus sequence of each region. Rearrangements, represented by reads from the rearranged derivative as well as the corresponding non-rearranged allele were instantly recognizable from a particular pattern of five vertices in the de Bruijn graph (a mathematical method used in *de novo* assembly of (short) read sequences) component of Velvet. Exact coordinates and features of junction sequence (for example, microhomology or non-templated sequence) were derived from this, following aligning to the reference genome, as though they were split reads.

See Supplementary Table 3 for summary of somatic variants. Annotation was according to ENSEMBL version 58.

Single nucleotide polymorphism (SNP) array hybridization using the Affymetrix SNP6.0 platform was performed according to Affymetrix protocols. Allele-specific copy number analysis of tumours was performed using ASCAT (v2.1.1), to generate integral allele-specific copy number profiles for the tumour cells⁵¹ (Supplementary Tables 4 and 5). ASCAT was also applied to next-generation sequencing data directly with highly comparable results.

We sampled 12.5% of the breast cancers for validation of substitutions, indels and/or rearrangements in order to make an assessment of the positive predictive value of mutation calling (Supplementary Table 6).

Further details of these processing steps as well as processing of transcriptomic and miRNA data (Supplementary Tables 7 and 8) can be found in Supplementary Methods.

Identification of novel breast cancer genes. To identify recurrently mutated driver genes, a dN/dS method that considers the mutation spectrum, the sequence of each gene, the impact of coding substitutions (synonymous, missense, nonsense, splice site) and the variation of the mutation rate across genes^{52,53} was used for substitutions (Supplementary Table 9). Owing to the lack of a neutral reference for the indel rate in coding sequences, a different approach was required (Supplementary Table 10, Supplementary Methods for details). To detect genes under significant selective pressure by either point mutations or indels, for each gene, the *P* values from the dN/dS analysis of substitutions and from the recurrence analysis of indels were combined using Fisher's method. Multiple testing correction (Benjamini-Hochberg FDR) was performed separately for the more than 600 putative driver genes and for all other genes, stratifying the FDR correction to increase sensitivity (as described in ref. 54). To achieve a low false discovery rate, a conservative *q*-value cutoff of <0.01 was used to determine statistical significance (Supplementary Table 11).

This analysis was applied to the new whole-genome sequences of 560 breast cancers as well as a further 772 breast cancers that have been sequenced previously by other institutions.

See Supplementary Methods for detailed explanations of these methods.

Recurrence in the non-coding regions. *Partitioning the genome into functional regulatory elements/gene features.* To identify non-coding regions with significant recurrence, we used a method similar to the one described for searching for novel indel drivers (see Supplementary Methods for detailed description).

The genome was partitioned according to different sets of regulatory elements/gene features, with a separate analysis performed for each set of elements, including exons (*n* = 20,245 genes), core promoters (*n* = 20,245 genes, where a core promoter is the interval (−250, +250) bp from any transcription start site (TSS) of a coding transcript of the gene, excluding any overlap with coding regions), 5' UTR (*n* = 9,576 genes), 3' UTR (*n* = 19,502 genes), intronic regions flanking exons (*n* = 20,212 genes, represents any intronic sequence within 75 bp from an exon, excluding any base overlapping with any of the above elements), any other sequence within genes (*n* = 18,591 genes, for every protein-coding gene, this contains any region within the start and end of transcripts not included in any of the above categories), non-coding RNAs (ncRNAs) (*n* = 10,684, full length lincRNAs, miRNAs or rRNAs), enhancers⁵⁵ (*n* = 194,054), ultra-conserved regions (*n* = 187,057, a collection of regions under negative selection based on 1,000 genomes data²⁰).

Every element set listed above was analysed separately to allow for different mutation rates across element types and to stratify the FDR correction⁵⁴. Within each set of elements, we used a negative binomial regression approach to learn the underlying variation of the mutation rate across elements. The offset reflects the expected number of mutations in each element assuming uniform mutation rates across them (that is, $E_{\text{subs,element}} = \sum_{j \in \{1,2,\dots,192\}} (r_j S_j t)$, and, $E_{\text{indels,element}} = \mu_{\text{indel}} S_{\text{indel,element}}$) (see Supplementary Methods 7 for a detailed description and definition of all parameters). As covariate here we used the local density of mutations in neighbouring non-coding regions, corrected for sequence composition and trinucleotide mutation rates (that is, the *t* parameter of the dN/dS equations described in section 7.1 of Supplementary Methods). Normalized local rates were pre-calculated for 100 kb non-overlapping bins of the genome and used in all analyses. Other covariates (expression, replication time or Hi-C (genome-wide chromosome conformation capture)) were not used here as they were not found to substantially improve the model once the local mutation rate was used as a covariate. A separate regression analysis was performed for substitutions and indels, to account for the different level of uncertainty in the distribution of substitution and indel rates across elements.

$$\text{model}_{\text{subs}} = \text{glm.nb}(\text{formula} = n_{\text{subs}} \sim \text{offset}(\log(E_{\text{subs}})) + \mu_{\text{local,subs}})$$

$$\text{model}_{\text{indels}} = \text{glm.nb}(\text{formula} = n_{\text{indels}} \sim \text{offset}(\log(E_{\text{indels}})) + \mu_{\text{local,indels}})$$

The observed counts for each element ($n_{\text{subs,element}}$ and $n_{\text{indels,element}}$) are compared to the background distributions using a negative binomial test, with the estimated overdispersion parameters (θ_{subs} and θ_{indels}) estimated by the negative binomial regression, yielding *P* values for substitution and indel recurrence for each element. These *P* values were combined using Fisher's method and corrected for multiple testing using FDR (Supplementary Table 16a).

Partitioning the genome into discrete bins. We performed a genome-wide screening of recurrence in 1 kb non-overlapping bins. We employed the method described in earlier section, using as covariate the local mutation rate calculated from 5 Mb up and downstream from the bin of interest and excluding any low-coverage region from the estimate (Supplementary Table 16b, Extended Data Fig. 3a for example). Significant hits were subjected to manual curation to remove false positives caused by sequencing or mapping artefacts.

Mutational signatures analysis. Mutational signatures analysis was performed following a three-step process: (i) hierarchical *de novo* extraction based on somatic substitutions and their immediate sequence context, (ii) updating the set of consensus signatures using the mutational signatures extracted from breast cancer genomes, and (iii) evaluating the contributions of each of the updated consensus signatures in each of the breast cancer samples. These three steps are discussed in more details in the next sections.

Hierarchical de novo extraction of mutational signatures. The mutational catalogues of the 560 breast cancer whole genome sequences were analysed for mutational signatures using a hierarchical version of the Wellcome Trust Sanger Institute mutational signatures framework²⁵. Briefly, we converted all mutation data into a matrix, *M*, that is made up of 96 features comprising mutations counts for each mutation type (C>A, C>G, C>T, T>A, T>C, and T>G; all substitutions are referred to by the pyrimidine of the mutated Watson-Crick base pair) using each possible 5' (C, A, G, and T) and 3' (C, A, G, and T) context for all samples. After conversion, the previously developed algorithm was applied in a hierarchical manner to the matrix *M* that contains *K* mutation types and *G* samples. The algorithm deciphers the minimal set of mutational signatures that optimally explains the proportion of each mutation type and then estimates the contribution of each signature across the samples. More specifically, the algorithm makes use of a

well-known blind source separation technique, termed non-negative matrix factorization (NNMF). NNMF identifies the matrix of mutational signature, P , and the matrix of the exposures of these signatures, E , by minimizing a Frobenius norm, while maintaining non-negativity:

$$\min_{P \in \mathbb{M}_{R+}^{(K, N)}, E \in \mathbb{M}_{R+}^{(N, G)}} \|M - PE\|_F^2$$

K is the number of mutation types (that is, 96), and \tilde{K} is the number of mutation types after dimensionality reduction. $P \in \mathbb{M}_{R+}^{(K, N)}$ is a matrix of real non-negative numbers of dimension $\tilde{K} \times N$. $E \in \mathbb{M}_{R+}^{(N, G)}$ is a matrix of real non-negative numbers of dimension $N \times G$. The method for deciphering mutational signatures, including evaluation with simulated data and list of limitations, can be found in ref. 25. The framework was applied in a hierarchical manner to increase its ability to find mutational signatures present in few samples as well as mutational signatures exhibiting a low mutational burden. More specifically, after application to the original matrix M containing 560 samples, we evaluated the accuracy of explaining the mutational patterns of each of the 560 breast cancers with the extracted mutational signatures. All samples that were well-explained by the extracted mutational signatures were removed and the framework was applied to the remaining sub-matrix of M . This procedure was repeated until the extraction process did not reveal any new mutational signatures. Overall, the approach extracted 12 unique mutational signatures operative across the 560 breast cancers (Fig. 3, Supplementary Table 21). **Updating the set of consensus mutational signatures.** The 12 hierarchically extracted breast cancer signatures were compared to the census of consensus mutational signatures²⁵. Of the 12 signatures, 11 closely resembled previously identified mutational patterns. The patterns of these 11 signatures, weighted by the numbers of mutations contributed by each signature in the breast cancer data, were used to update the set of consensus mutational signatures as previously performed in ref. 25. One of the 12 extracted signatures is novel and at present, unique for breast cancer. This novel signature is consensus signature 30 (<http://cancer.sanger.ac.uk/cosmic/signatures>).

Evaluating the contributions of consensus mutational signatures in 560 breast cancers. The complete compendium of consensus mutational signatures that was found in breast cancer includes: signatures 1, 2, 3, 5, 6, 8, 13, 17, 18, 20, 26, and 30. We evaluated the presence of all of these signatures in the 560 breast cancer genomes by re-introducing them into each sample. More specifically, the updated set of consensus mutational signatures was used to minimize the constrained linear function for each sample:

$$\min_{e_i \geq 0} \|\mathbf{m} - \sum_{i=1}^N (\mathbf{p}_i e_i)\|_F^2$$

Here, \mathbf{m} is a vector with 96 components corresponding to the counts of each of the mutation types in a sample, \mathbf{p}_i represents a vector with 96 components (corresponding to a consensus mutational signature i), e_i is a non-negative scalar reflecting the number of mutations contributed by signature i in that sample. N is equal to 12 and it reflects the number of all possible signatures that can be found in a single breast cancer sample. Mutational signatures that did not contribute large numbers (or proportions) of mutations or that did not significantly improve the correlation between the original mutational pattern of the sample and the one generated by the mutational signatures were excluded from the sample. This procedure reduced over-fitting the data and allowed only the essential mutational signatures to be present in each sample (Supplementary Table 21b).

Kataegis. Kataegis, or foci of localized hypermutation, has been previously defined²⁵ as 6 or more consecutive mutations with an average intermutation distance of less than or equal to 1,000 bp. Kataegis were sought in 560 whole-genome sequenced breast cancers from high-quality base substitution data using the method described previously²⁵. This method likely misses some foci of kataegis sacrificing sensitivity of detection for a higher positive predictive value of kataegis foci (Supplementary Table 21c).

Rearrangement signatures. *Clustered vs non-clustered rearrangements.* We sought to separate rearrangements that occurred as focal catastrophic events or focal driver amplicons from genome-wide rearrangement mutagenesis using a piecewise constant fitting method. For each sample, both breakpoints of each rearrangement were considered individually and all breakpoints were ordered by chromosomal position. The inter-rearrangement distance, defined as the number of base pairs from one rearrangement breakpoint to the one immediately preceding it in the reference genome, was calculated. Putative regions of clustered rearrangements were identified as having an average inter-rearrangement distance that was at least 10 times greater than the whole-genome average for the individual sample. Piecewise constant fitting parameters used were $\gamma = 25$ and $k_{\min} = 10$, with γ as the parameter that controls smoothness of segmentation, and k_{\min} the minimum number of breakpoints in a segment.

The respective partner breakpoint of all breakpoints involved in a clustered region are likely to have arisen at the same mechanistic instant and so were considered as being involved in the cluster even if located at a distant chromosomal site. The rearrangements within clusters ('clustered') and not within clusters ('non-clustered') are summarized in Extended Data Table 4.

Classification: types and size. In both classes of rearrangements, clustered and non-clustered, rearrangements were subclassified into deletions, inversions and tandem duplications, and then further subclassified according to size of the rearranged segment (1–10 kb, 10–100 kb, 100 kb–1 Mb, 1–10 Mb, more than 10 Mb). The final category in both groups was interchromosomal translocations.

Rearrangement signatures by NNMF. The classification produces a matrix of 32 distinct categories of structural variants across 544 breast cancer genomes. This matrix was decomposed using the previously developed approach for deciphering mutational signatures by searching for the optimal number of mutational signatures that best explains the data without over-fitting the data²⁵ (Supplementary Table 21d, e).

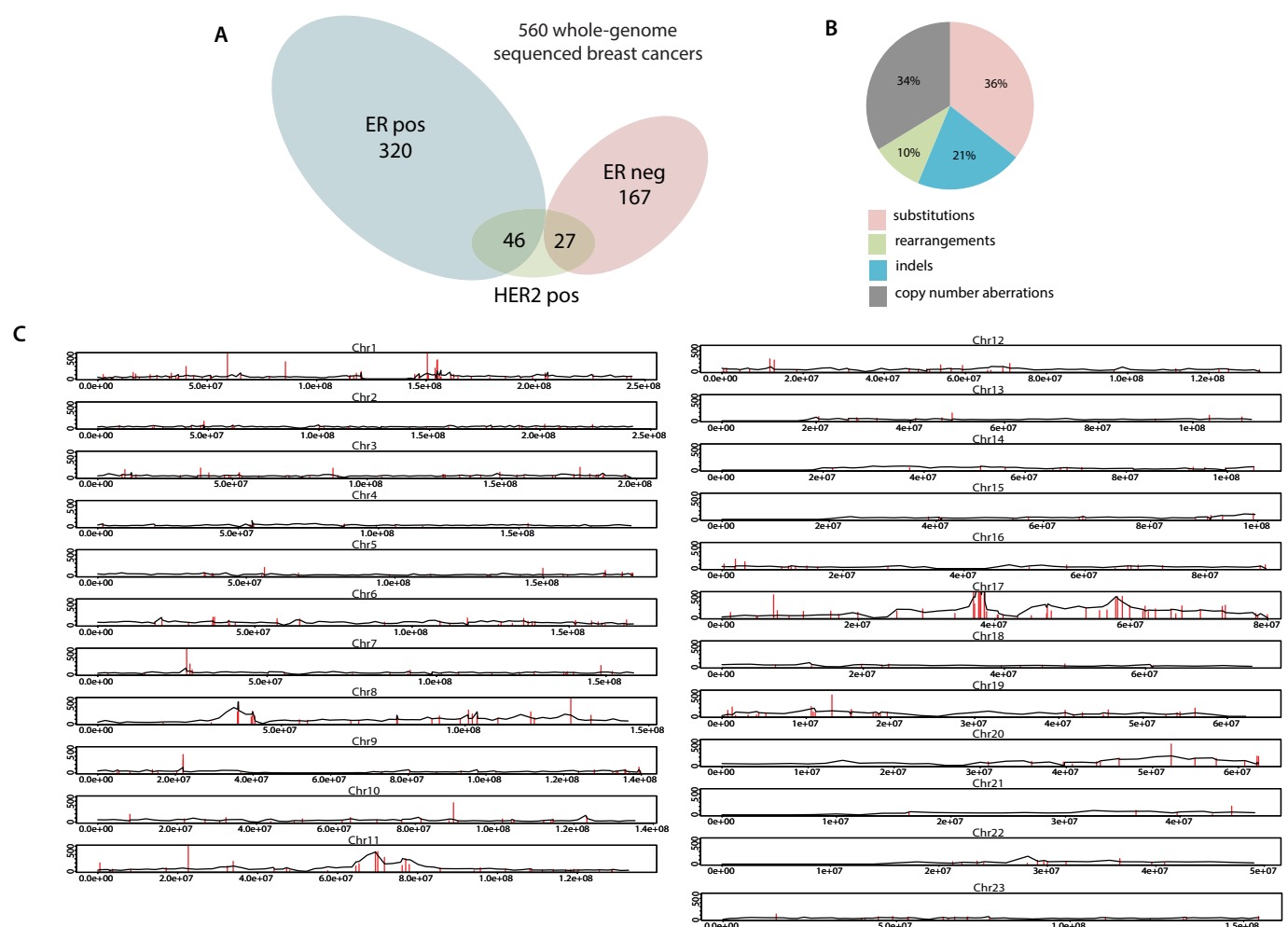
Consensus clustering of rearrangement signatures. To identify subgroups of samples sharing similar combinations of six identified rearrangement signatures derived from whole genome sequencing analysis we performed consensus clustering using the ConsensusClusterPlus R package⁵⁶. Input data for each sample ($n = 544$, a subset of the full sample cohort) was the proportion of rearrangements assigned to each of the six signatures. Thus, each sample has 6 data values, with a total sum of 1. Proportions for each signature were mean-centred across samples before clustering. The following settings were used in the consensus clustering: number of repetitions = 1000; pItem = 0.9 (resampling frequency samples); pFeature = 0.9 (resampling frequency); Pearson distance metric; Ward linkage method.

Distribution of mutational signatures relative to genomic architecture. Following extraction of mutational signatures and quantification of the exposures (or contributions) of each signature to each sample, a probability for each mutation belonging to each mutation signature (for a given class of mutation for example, substitutions) was assigned⁴².

The distribution of mutations as signatures were assessed across multiple genomic features including replication time, strands, transcriptional strands and nucleosome occupancy. See ref. 42 for technical details, per signature results.

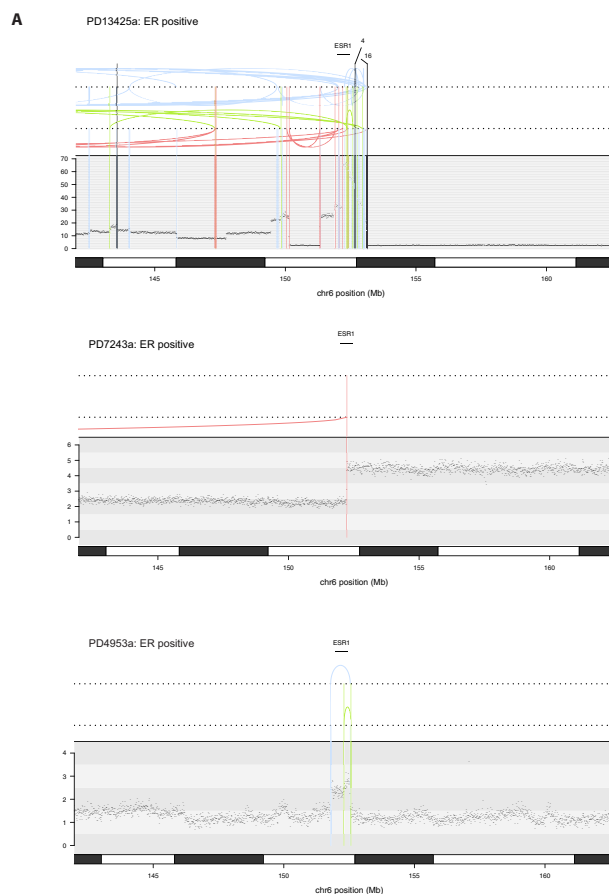
Individual patient whole-genome profiles. Breast cancer whole-genome profiles were adapted from the R Circos package⁵⁷. See <http://cancer.sanger.ac.uk/cosmic/sample/genomes> for individual patient genome profiles. Features depicted in circos plots from outermost rings heading inwards: Karyotypic ideogram outermost. Base substitutions next, plotted as rainfall plots (log₁₀ intermutation distance on radial axis, dot colours: blue, C>A; black, C>G; red, C>T; grey, T>A; green, T>C; pink, T>G). Ring with short green lines, insertions; ring with short red lines, deletions. Major copy number allele (green, gain) ring, minor copy number allele ring (pink, loss). Central lines represent rearrangements (green, tandem duplications; pink, deletions; blue, inversions; grey, interchromosomal events). In each profile, the top right-hand panel displays the number of mutations contributing to each mutation signature extracted using NNMF in individual cancers. Middle right-hand panel represents indels. Bottom right corner shows histogram of rearrangements present in this cancer. Bottom left corner shows all curated driver mutations, top- and middle-left panels show clinical and pathology data respectively.

47. Kozarewa, I. *et al.* Amplification-free Illumina sequencing-library preparation facilitates improved mapping and assembly of (G+C)-biased genomes. *Nature Methods* **6**, 291–295 (2009).
48. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
49. Ye, K., Schulz, M. H., Long, Q., Apweiler, R. & Ning, Z. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics* **25**, 2865–2871 (2009).
50. Zerbino, D. R. & Birney, E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res.* **18**, 821–829 (2008).
51. Van Loo, P. *et al.* Allele-specific copy number analysis of tumors. *Proc. Natl Acad. Sci. USA* **107**, 16910–16915 (2010).
52. Greenman, C., Wooster, R., Futreal, P. A., Stratton, M. R. & Easton, D. F. Statistical analysis of pathogenicity of somatic mutations in cancer. *Genetics* **173**, 2187–2198 (2006).
53. Lawrence, M. S. *et al.* Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **499**, 214–218 (2013).
54. Sun, L., Craiu, R. V., Paterson, A. D. & Bull, S. B. Stratified false discovery control for large-scale hypothesis testing with application to genome-wide association studies. *Genet. Epidemiol.* **30**, 519–530 (2006).
55. The ENCODE Project Consortium An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
56. Wilkerson, M. D. & Hayes, D. N. ConsensusClusterPlus: a class discovery tool with confidence assessments and item tracking. *Bioinformatics* **26**, 1572–1573 (2010).
57. Zhang, H., Meltzer, P. & Davis, S. RCircos: an R package for Circos 2D track plots. *BMC Bioinformatics* **14**, 244 (2013).

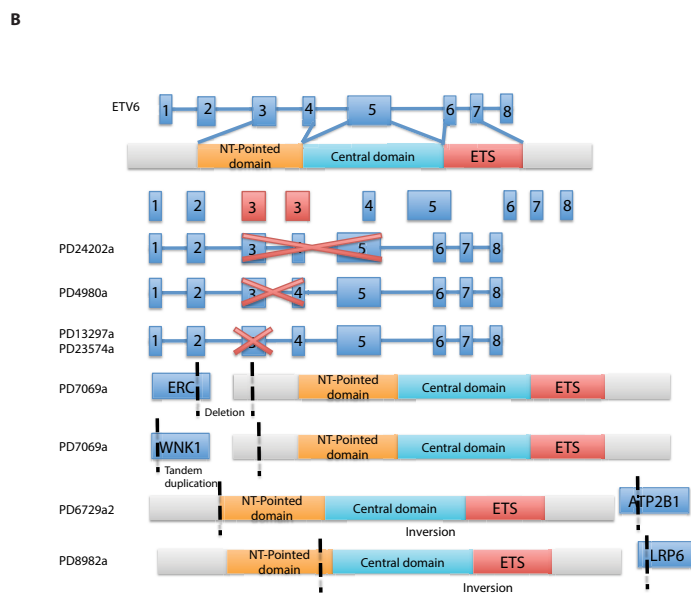


Extended Data Figure 1 | Landscape of driver mutations. **a**, Summary of subtypes of cohort of 560 breast cancers. **b**, Driver mutations by mutation type. **c**, Distribution of rearrangements throughout the genome. Black

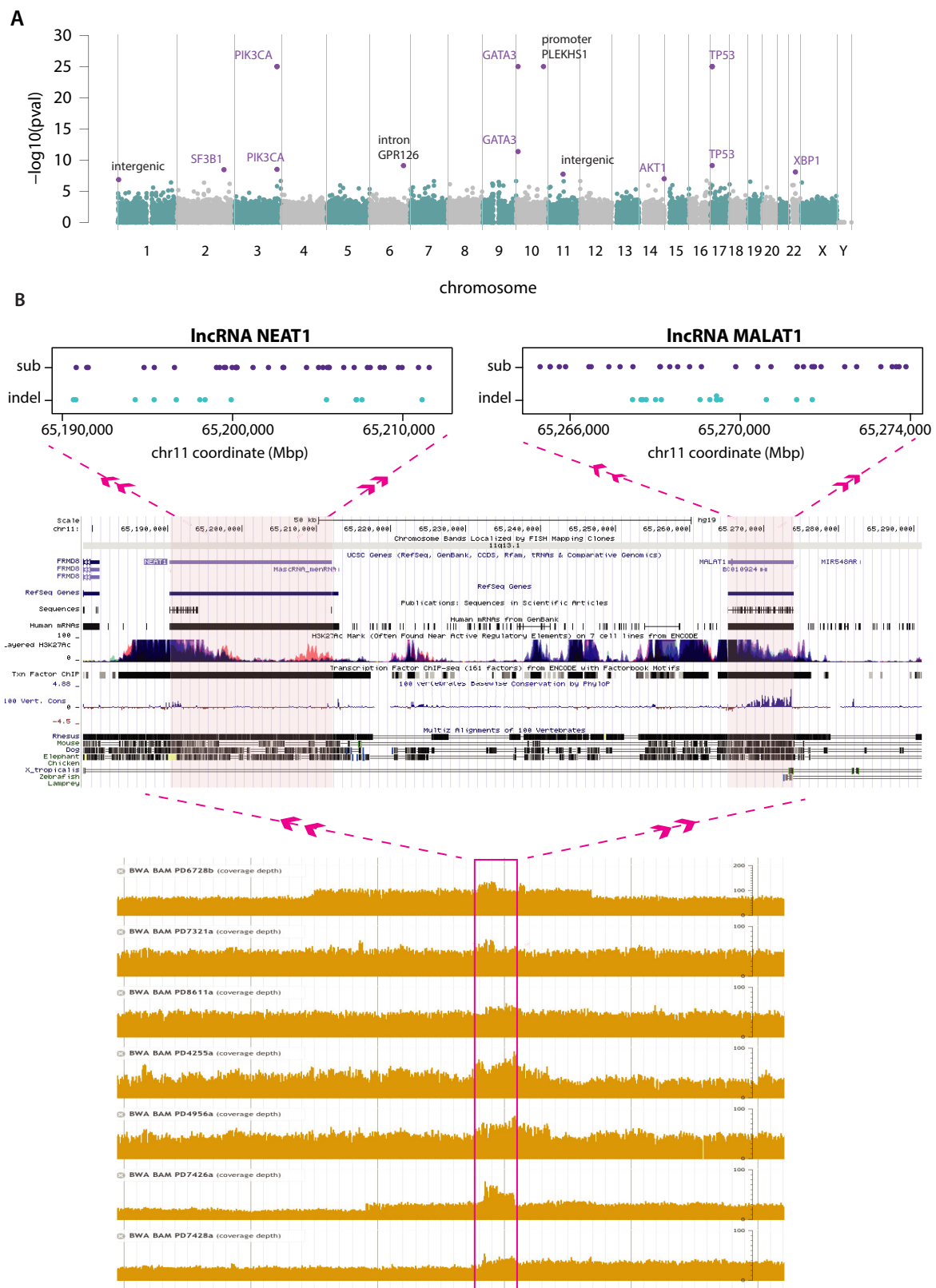
line represents background rearrangement density (calculation based on rearrangement breakpoints in intergenic regions only). Red lines represent frequency of rearrangement within breast cancer genes.



Extended Data Figure 2 | Rearrangements in oncogenes. a, Variation in rearrangement and copy number events affecting *ESR1*. Clear amplification in top panel, translocation of *ESR1* in middle panel and focused tandem duplication events in bottom panel. **b**, Predicted outcomes of some rearrangements affecting *ETV6*. Red crosses indicate exons

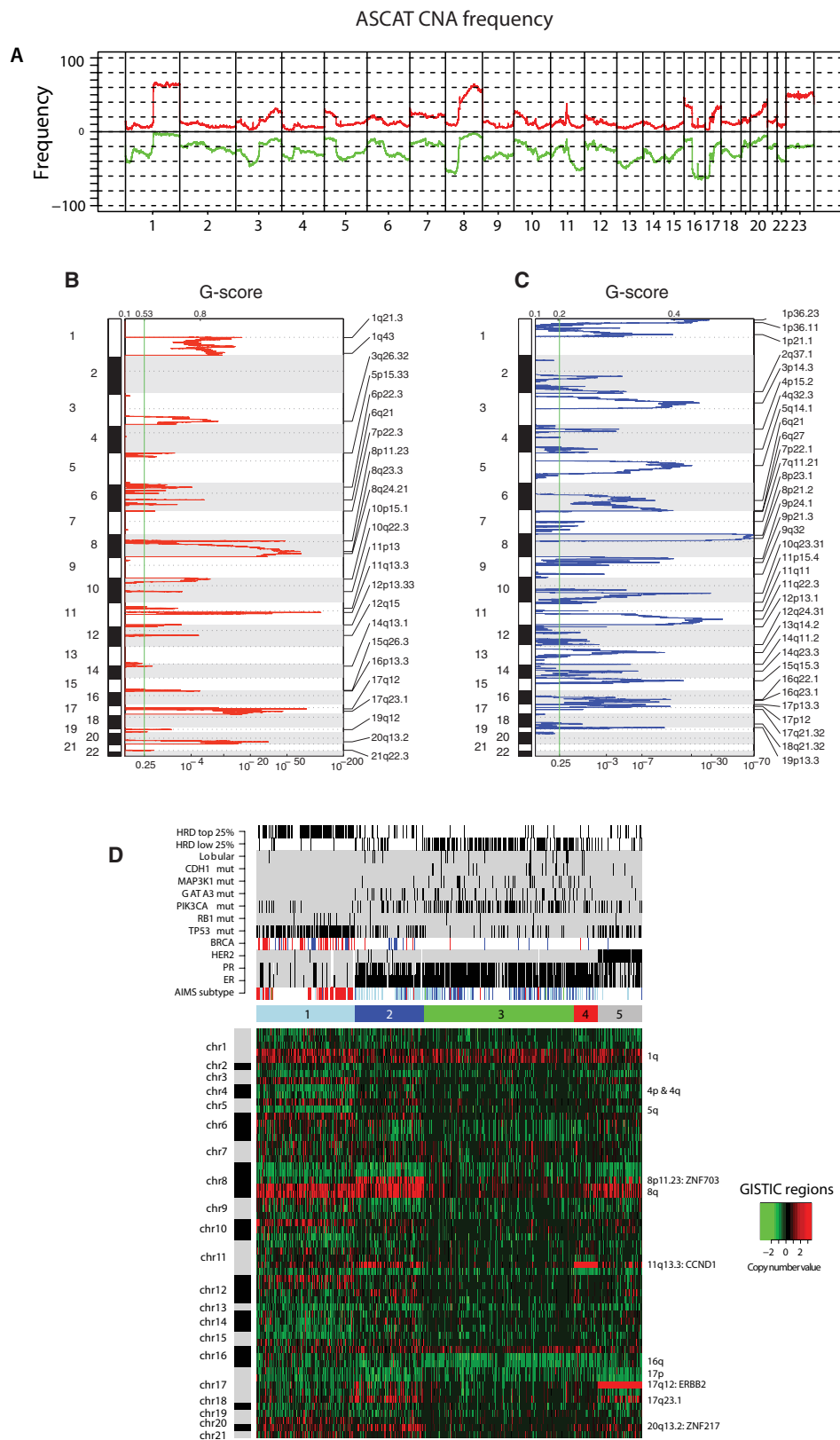


deleted as a result of rearrangements within the *ETV6* genes, black dotted lines indicate rearrangement break points resulting in fusions between *ETV6* and *ERC*, *WNK1*, *ATP2B1* or *LRP6*. *ETV6* domains indicated are: N-terminal (NT) pointed domain and E26 transformation-specific DNA binding domain (ETS).



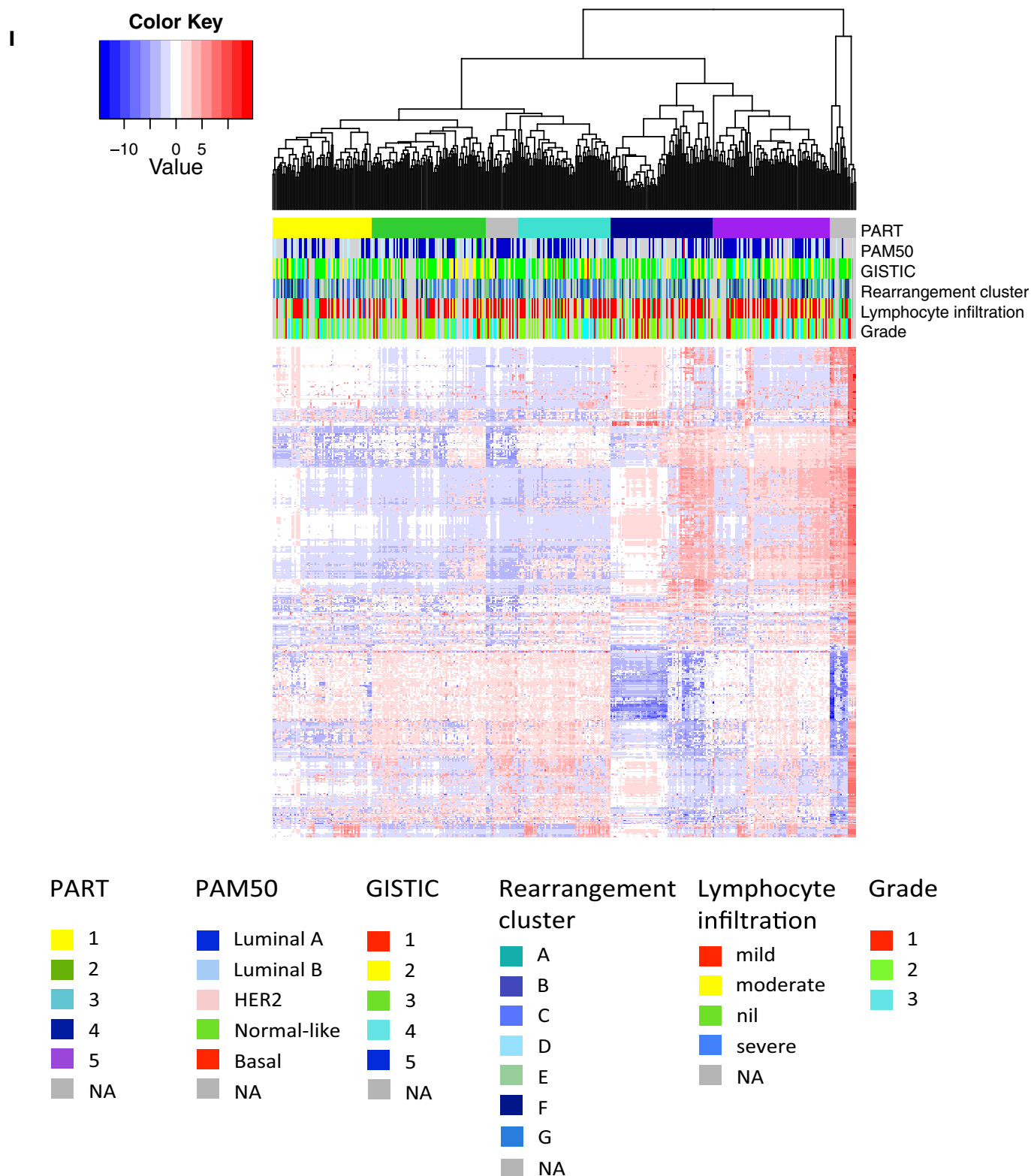
Extended Data Figure 3 | Recurrent non-coding events in breast cancers. **a**, Manhattan plot demonstrating sites with most significant P values as identified by binning analysis. Purple highlighted sites were also detected by the method seeking recurrence when partitioned by genomic features. **b**, Locus at chr11 65 Mb, which was identified by independent analyses as being more mutated than expected by chance.

Bottom, a rearrangement hotspot analysis identified this region as a tandem duplication hotspot, with nested tandem duplications noted at this site. Partitioning the genome into different regulatory elements, an analysis of substitutions and indels identified lncRNAs *MALAT1* and *NEAT1* (topmost panels) with significant P values.



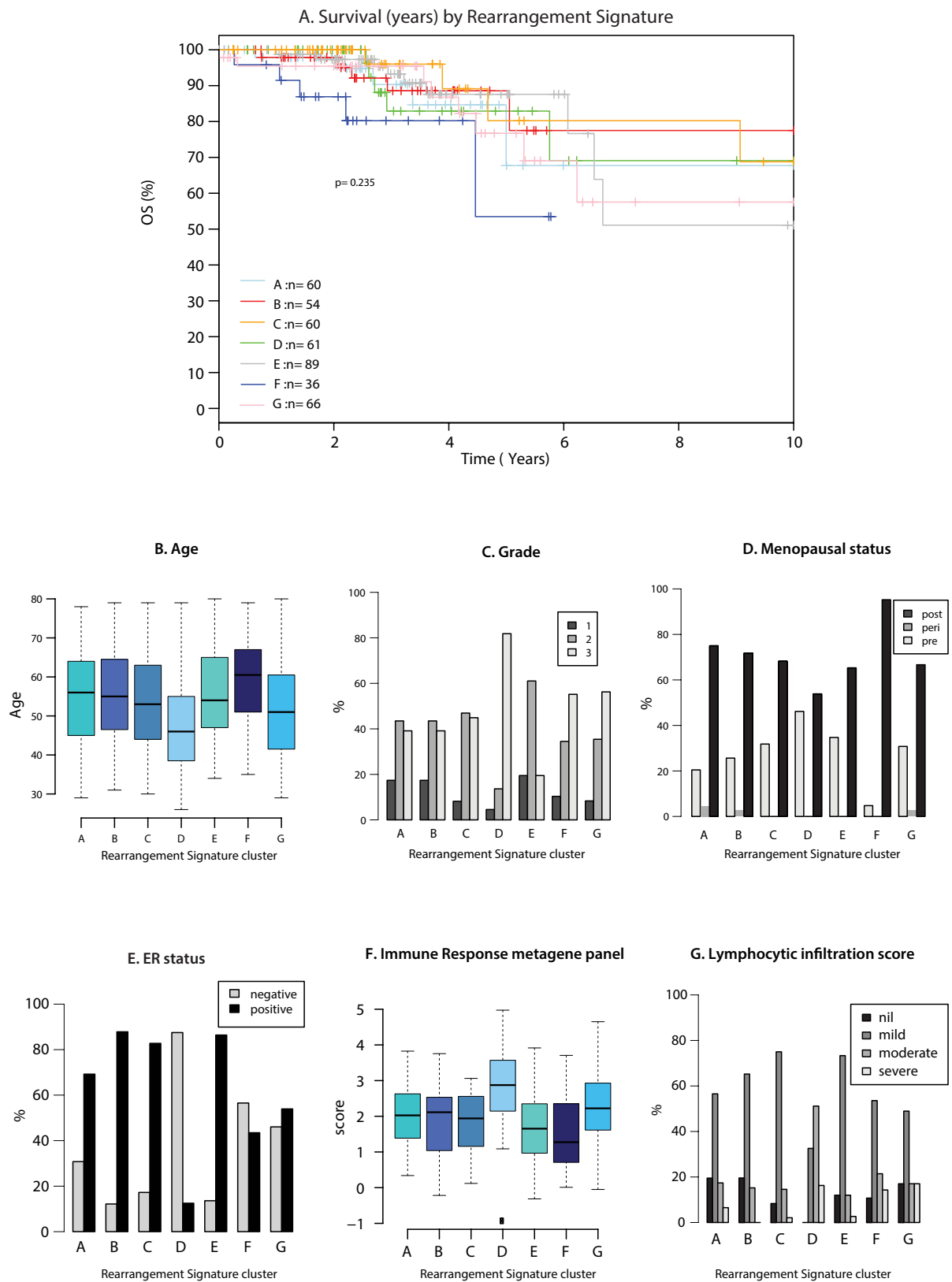
Extended Data Figure 4 | Copy number analyses. **a**, Frequency of copy number aberrations across the cohort. Chromosome position along *x* axis, frequency of copy number gains (red) and losses (green) *y* axis. **b**, Identification of focal recurrent copy number gains by the GISTIC method (Supplementary Methods). **c**, Identification of focal recurrent copy number losses by the GISTIC method. **d**, Heatmap of GISTIC regions following unsupervised hierarchical clustering. Five cluster

groups are noted and relationships with expression subtype (basal, red; luminal B, light blue; luminal A, dark blue), immunohistochemistry status (ER, PR, HER2 status; black, positive), abrogation of *BRCA1* (red) and *BRCA2* (blue) (whether germline, somatic or through promoter hypermethylation), driver mutations (black, positive), HRD index (top 25% or lowest 25%; black, positive).

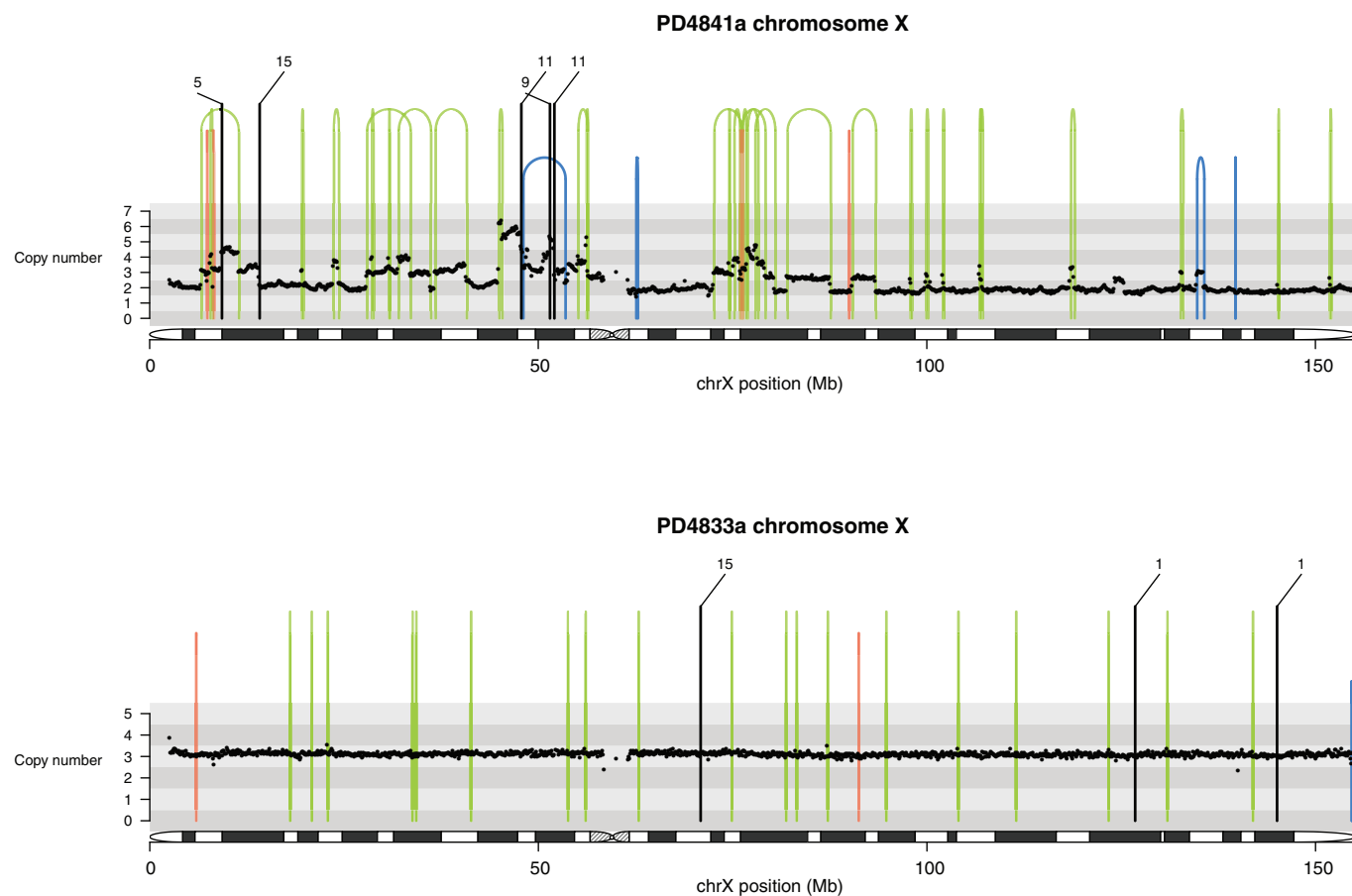


Extended Data Figure 5 | miRNA analyses. Hierarchical clustering of the most variant miRNAs using complete linkage and Euclidean distance. miRNA clusters were assigned using the partitioning algorithm using recursive thresholding (PART) method. Five main patient clusters were revealed. The horizontal annotation bars show (from top to bottom):

PART cluster group, PAM50 mRNA expression subtype, GISTIC cluster, rearrangement cluster, lymphocyte infiltration score and histological grade. The heatmap shows clustered and centred miRNA expression data (log₂ transformed). Details on colour coding of the annotation bars are presented below the heatmap.

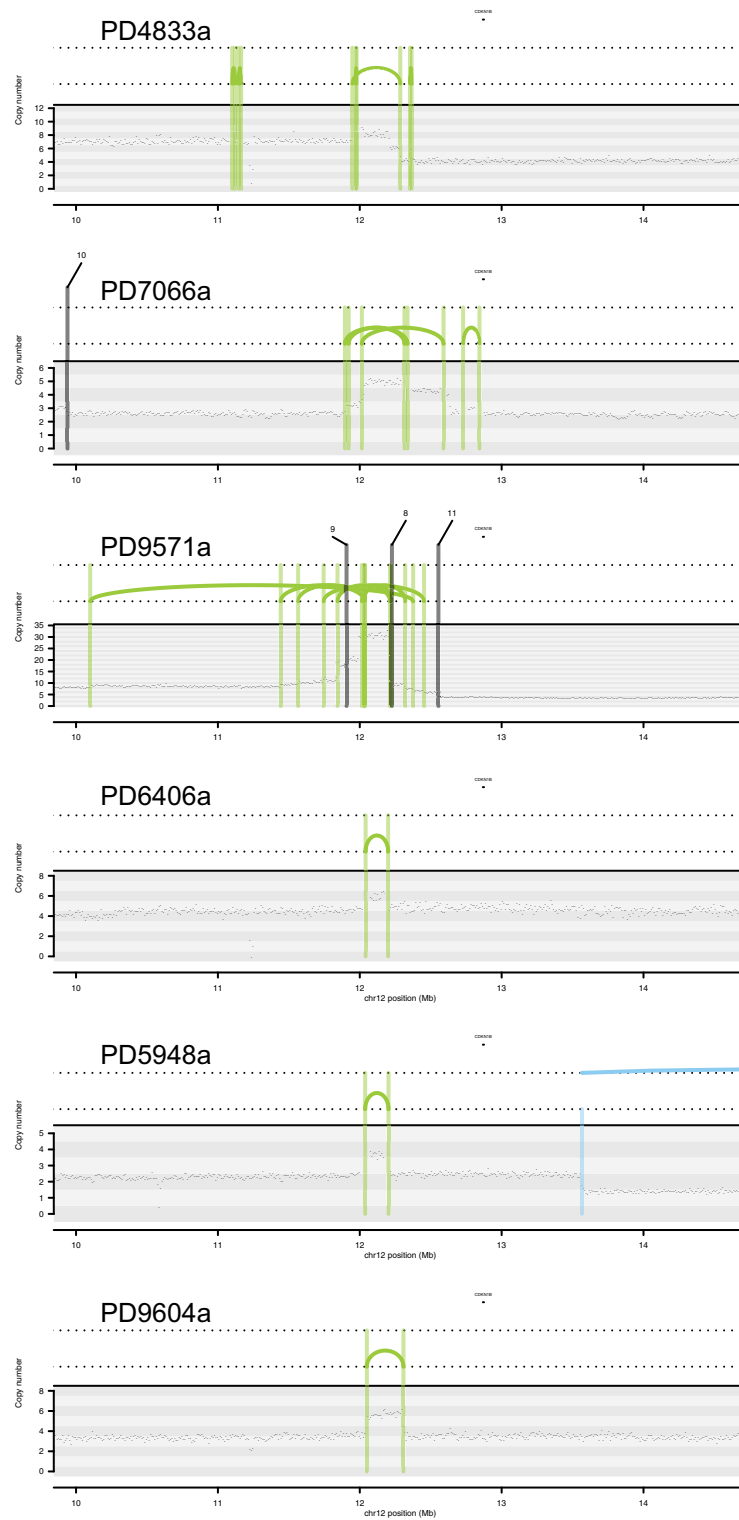


Extended Data Figure 6 | Rearrangement cluster groups and associated features. a, Overall survival (OS) by rearrangement cluster group. **b,** Age of diagnosis. **c,** Tumour grade. **d,** Menopausal status. **e,** ER status. **f,** Immune response metagene panel. **g,** Lymphocytic infiltration score.

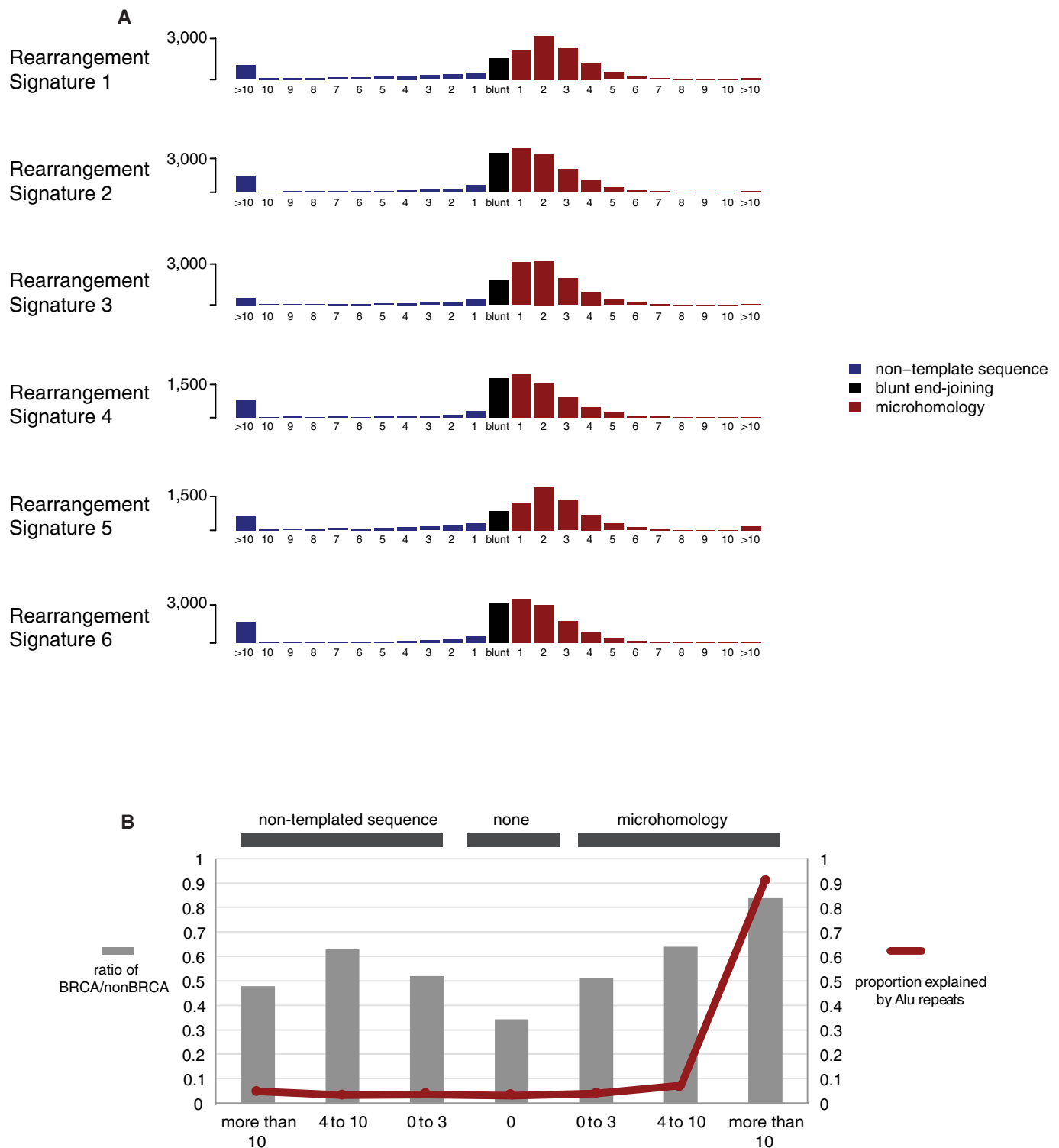


Extended Data Figure 7 | Contrasting tandem duplication phenotypes. Contrasting tandem duplication phenotypes of two breast cancers using chromosome X. Copy number (y axis) depicted as black dots. Lines represent rearrangements breakpoints (green, tandem duplications; pink, deletions; blue, inversions; black, translocations with partner

breakpoint provided). Top, PD4841a has numerous large tandem duplications (>100 kb, rearrangement signature 1), whereas PD4833a has many short tandem duplications (<10 kb, rearrangement signature 3) appearing as 'single' lines in its plot.

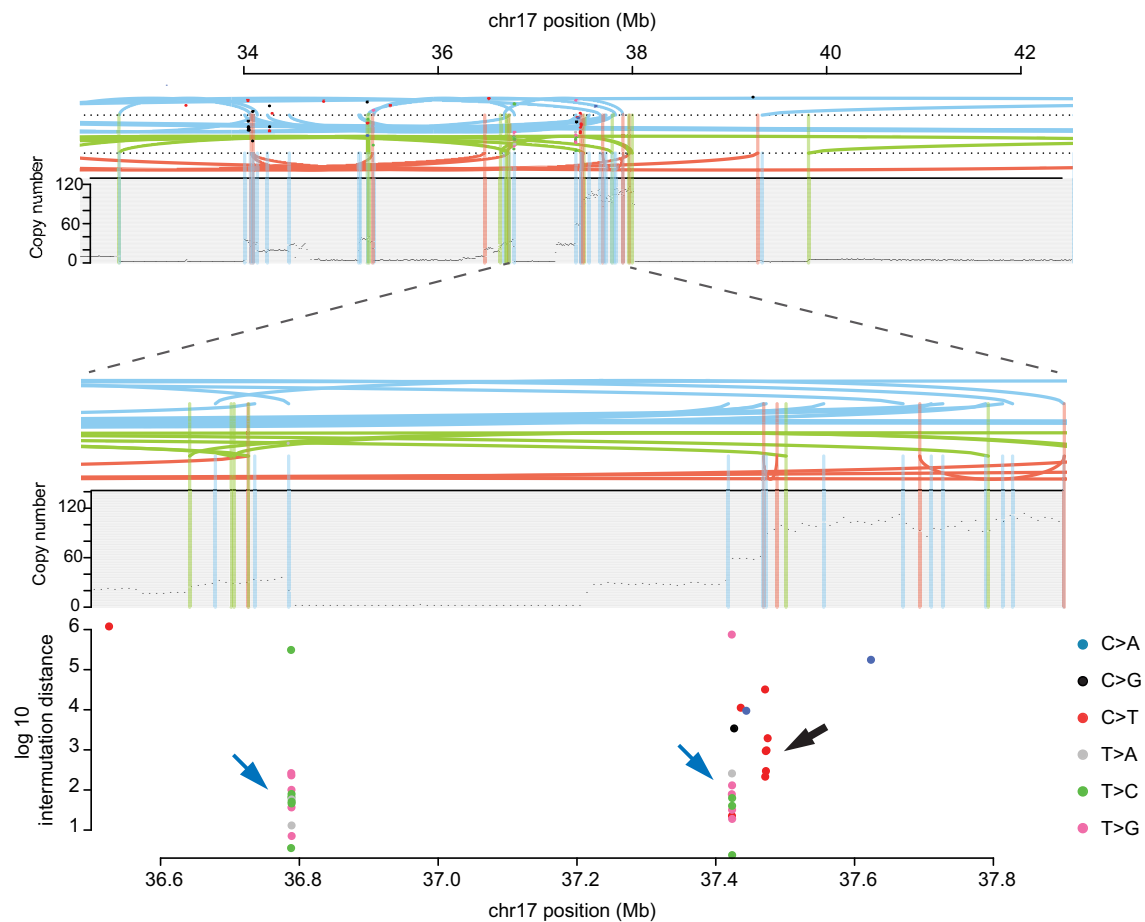


Extended Data Figure 8 | Hotspots of tandem duplications. A tandem duplication hotspot occurring in six different patients.

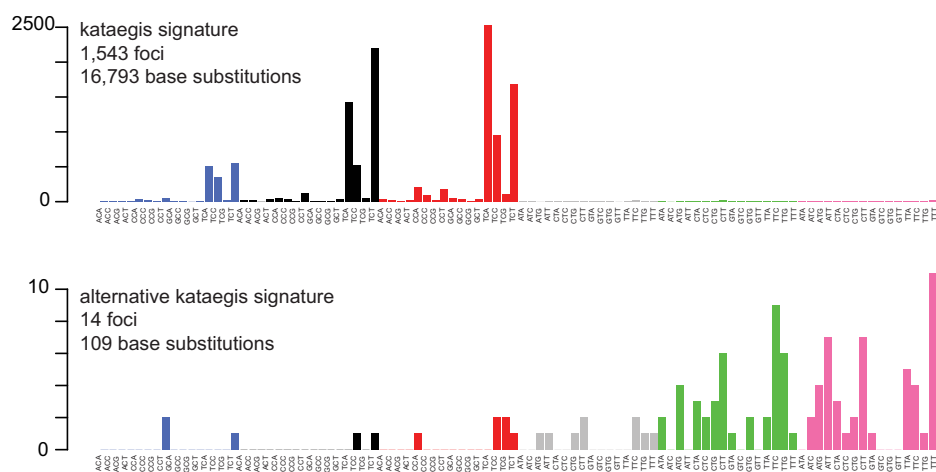


Extended Data Figure 9 | Rearrangement breakpoint junctions. **a**, Breakpoint features of rearrangements in 560 breast cancers by rearrangement signature. **b**, Breakpoint features in BRCA and non-BRCA cancers.

A



B



Extended Data Figure 10 | Signatures of focal hypermutation.

a, Kataegis and alternative kataegis occurring at the same locus (ERBB2 amplicon in PD13164a). Copy number (y axis) depicted as black dots. Lines represent rearrangements breakpoints (green, tandem duplications; pink, deletions; blue, inversions). Top, an ~10 Mb region including the ERBB2 locus. Middle, zoomed-in tenfold to an ~1 Mb window

highlighting co-occurrence of rearrangement breakpoints, with copy number changes and three different kataegis loci. Bottom, demonstrates kataegis loci in more detail. log₁₀ intermutation distance on y axis. Black arrow, kataegis; blue arrows, alternative kataegis. **b**, Sequence context of kataegis and alternative kataegis identified in this data set.

Proteogenomics connects somatic mutations to signalling in breast cancer

Philipp Mertins^{1*}, D. R. Mani^{1*}, Kelly V. Ruggles^{2*}, Michael A. Gillette^{1,3*}, Karl R. Clauser¹, Pei Wang⁴, Xianlong Wang⁵, Jana W. Qiao¹, Song Cao⁶, Francesca Petralia⁴, Emily Kawaler², Filip Mundt^{1,7}, Karsten Krug¹, Zhidong Tu⁴, Jonathan T. Lei⁸, Michael L. Gatz⁹, Matthew Wilkerson⁹, Charles M. Perou⁹, Venkata Yellapantula⁶, Kuan-lin Huang⁶, Chenwei Lin⁵, Michael D. McLellan⁶, Ping Yan⁵, Sherri R. Davies¹⁰, R. Reid Townsend¹⁰, Steven J. Skates¹¹, Jing Wang¹², Bing Zhang¹², Christopher R. Kinsinger¹³, Mehdi Mesri¹³, Henry Rodriguez¹³, Li Ding⁶, Amanda G. Paulovich⁵, David Fenyo², Matthew J. Ellis⁸, Steven A. Carr¹ & the NCI CPTAC†

Somatic mutations have been extensively characterized in breast cancer, but the effects of these genetic alterations on the proteomic landscape remain poorly understood. Here we describe quantitative mass-spectrometry-based proteomic and phosphoproteomic analyses of 105 genomically annotated breast cancers, of which 77 provided high-quality data. Integrated analyses provided insights into the somatic cancer genome including the consequences of chromosomal loss, such as the 5q deletion characteristic of basal-like breast cancer. Interrogation of the 5q *trans*-effects against the Library of Integrated Network-based Cellular Signatures, connected loss of *CETN3* and *SKP1* to elevated expression of epidermal growth factor receptor (EGFR), and *SKP1* loss also to increased SRC tyrosine kinase. Global proteomic data confirmed a stromal-enriched group of proteins in addition to basal and luminal clusters, and pathway analysis of the phosphoproteome identified a G-protein-coupled receptor cluster that was not readily identified at the mRNA level. In addition to ERBB2, other amplicon-associated highly phosphorylated kinases were identified, including CDK12, PAK1, PTK2, RIPK2 and TLK2. We demonstrate that proteogenomic analysis of breast cancer elucidates the functional consequences of somatic mutations, narrows candidate nominations for driver genes within large deletions and amplified regions, and identifies therapeutic targets.

A central deficiency in our knowledge of cancer concerns how genomic changes drive the proteome and phosphoproteome to execute phenotypic characteristics^{1–4}. The initial proteomic characterization in the The Cancer Genome Atlas (TCGA) breast cancer study was performed using reverse phase protein arrays (RPPA); however this approach is restricted by antibody availability. To provide greater analytical breadth, the NCI Clinical Proteomic Tumor Analysis Consortium (CPTAC) is using mass spectrometry to analyse the proteomes of genome-annotated TCGA tumour samples^{5,6}. Here we describe integrated proteogenomic analyses of TCGA breast cancer samples representing the four principal mRNA-defined breast cancer intrinsic subtypes^{7,8}.

Proteogenomic analysis of TCGA samples

105 breast tumours previously characterized by the TCGA were selected for proteomic analysis after histopathological documentation (Supplementary Tables 1 and 2). The cohort included a balanced representation of PAM50-defined intrinsic subtypes⁹ including 25 basal-like, 29 luminal A, 33 luminal B, and 18 HER2 (*ERBB2*)-enriched tumours, along with 3 normal breast tissue samples. Samples were analysed by high-resolution accurate-mass tandem mass spectrometry (MS/MS) that included extensive peptide fractionation and

phosphopeptide enrichment (Extended Data Fig. 1a). An isobaric peptide labelling approach (iTRAQ) was employed to quantify protein and phosphosite levels across samples, with 37 iTRAQ 4-plexes analysed in total. A total of 15,369 proteins (12,405 genes) and 62,679 phosphosites were confidently identified with 11,632 proteins per tumour and 26,310 phosphosites per tumour on average (Supplementary Tables 3, 4 and Supplementary Methods). After filtering for observation in at least a quarter of the samples (Supplementary Methods, Extended Data Fig. 1b), 12,553 proteins (10,062 genes) and 33,239 phosphosites, with their relative abundances quantified across tumours, were used in subsequent analyses in this study. Stable longitudinal performance and low technical noise were demonstrated by repeated interspersed analyses of a single batch of patient-derived luminal and basal breast cancer xenograft samples¹⁰ (Extended Data Fig. 1d, e). Owing to the heterogeneous nature of breast tumours^{11–13}, and because proteomic analyses were performed on tumour fragments that were different from those used in the genomic analyses, rigorous pre-specified sample and data quality control metrics were implemented^{14,15} (Supplementary Discussion and Extended Data Figs 2, 3). Extensive analyses concluded that 28 of the 105 samples were compromised by protein degradation. These samples were excluded from further analysis with subsequent informatics focused on the 77 tumour samples and three biological replicates.

¹The Broad Institute of MIT and Harvard, Cambridge, Massachusetts 02142, USA. ²Department of Biochemistry and Molecular Pharmacology, New York University Langone Medical Center, New York, New York 10016, USA. ³Division of Pulmonary and Critical Care Medicine, Massachusetts General Hospital, Boston, Massachusetts 02114, USA. ⁴Department of Genetics and Genomic Sciences, Icahn Institute of Genomics and Multiscale Biology, Icahn School of Medicine at Mount Sinai New York, New York 10029, USA. ⁵Fred Hutchinson Cancer Research Center, Seattle, Washington 98109, USA. ⁶Department of Medicine, McDonnell Genome Institute, Siteman Cancer Center, Washington University School of Medicine, St. Louis, Missouri 63108, USA. ⁷Department of Oncology-Pathology, Karolinska Institute, 171 76 Stockholm, Sweden. ⁸Lester and Sue Smith Breast Center, Dan L. Duncan Comprehensive Cancer Center and Departments of Medicine and Molecular and Cellular Biology, Baylor College of Medicine, Houston, Texas 77030, USA. ⁹Department of Genetics, Lineberger Comprehensive Cancer Center, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599, USA. ¹⁰Department of Medicine, Washington University School of Medicine, St. Louis, Missouri 63110, USA. ¹¹Biostatistics Center, Massachusetts General Hospital Cancer Center, Boston, Massachusetts 02114, USA. ¹²Department of Biomedical Informatics and Department of Cancer Biology, Vanderbilt University School of Medicine, Nashville, Tennessee 37232, USA. ¹³National Cancer Institute, National Institutes of Health, Bethesda, Maryland 20892, USA.

*These authors contributed equally to this work.

†A list of participants and their affiliations appears in the Supplementary Information.



Figure 1 | Direct effects of genomic alterations on protein level.

a, b, Overlap of protein-coding single amino acid variants (**a**) and RNA splice junctions (**b**) not present in RefSeq v60 detected by DNA exome sequencing, RNA-seq, and LC-MS/MS. Proportions of novel variants are noted. **c**, Heat map of mutations/CNA and their effects on RNA and protein expression of breast-cancer-relevant genes across tumour and

normal samples. ER, PR, HER2 and PAM50 status are annotated. Median iTRAQ protein abundance ratio and the most frequently detected and differential phosphosite ratio are shown for each gene. Pearson correlations between MS/MS protein and RNA-seq, and MS/MS protein and RPPA are indicated.

Genome and transcriptomic variation was observed at the peptide level by searching MS/MS spectra not matched to RefSeq against a patient-specific sequence database (Fig. 1a). The database was constructed using the QUILTS software package¹⁶, leveraging RefSeq gene models based on whole-exome and RNA-seq data generated from portions of the same tumours and matched germline DNA (Fig. 1a, Supplementary Table 5). Although these analyses detected a number of single amino acid variants, frameshifts, and splice junctions, including splice isoforms that had been detected as only single transcript reads by RNA-seq (Fig. 1b, Supplementary Table 5), the number of genomic and transcriptomic variants that were confirmed as peptides by MS/MS was low (Supplementary Discussion). Sparse detection of

individual genomic variants by peptide sequencing has been noted in our previous studies¹⁶ and reflects limited coverage at the single amino acid level with current technology. However, quantitative MS/MS analysis of multiple peptides for each protein is used to reliably infer overall protein levels. This is an advantage of MS/MS, as antibody-based protein expression analysis is typically based on a single epitope. To illustrate this capability in the current data set, an initial analysis of three frequently mutated genes in breast cancer (*TP53*, *PIK3CA*, and *GATA3*) and three clinical biomarkers (oestrogen receptor (ER; *ESR1*), progesterone receptor (*PGR*), and *ERBB2*) was conducted (Fig. 1c, Supplementary Table 6, 7 and Supplementary Discussion). As expected, *TP53* missense mutations were associated with elevated

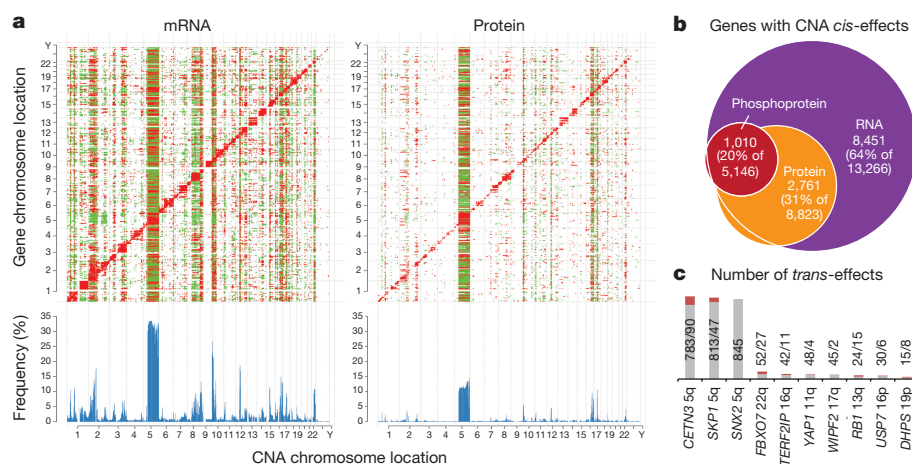


Figure 2 | Effects of CNAs on mRNA, protein, and phosphoprotein abundance. **a**, Correlations of CNA (x axes) to RNA and protein expression levels (y axes) highlight new CNA *cis*- and *trans*-effects. Significant ($FDR < 0.05$) positive (red) and negative (green) correlations between CNA and mRNAs or proteins are indicated. CNA *cis*-effects appear as a red diagonal line, CNA *trans*-effects as vertical stripes. Histograms show the fraction (%) of significant CNA *trans*-effects for

each CNA gene. **b**, Overlap of *cis*-effects observed at RNA, protein, and phosphoprotein levels ($FDR < 0.05$). **c**, *Trans*-effect regulatory candidates identified among those with significant protein *cis*-effects using LINCS CMap. Bars indicate total numbers of significant CNA–protein *trans*-effects (grey; $FDR < 0.05$) and overlap with regulated genes in LINCS knockdown profiles (red; 4 cell lines; moderated t -test $FDR < 0.1$).

MS/MS-based protein levels, as observed by RPPA, especially in basal-like breast cancer. *TP53* nonsense and frameshift mutations were associated with a decrease in *TP53* protein levels that was particularly pronounced in the MS/MS data. In contrast, the mostly C-terminal *GATA3* frameshift alterations did not result in decreased protein expression when measured by the median of all *GATA3* peptides, suggesting that these proteins are expressed despite truncation. No consistent effect of somatic *PIK3CA* mutation was observed at the level of protein expression. Good Pearson correlations between RNA-seq and MS/MS protein-expression levels were found for *ESR1* ($r = 0.74$), *PGR* ($r = 0.74$), *ERBB2* ($r = 0.84$) and *GATA3* ($r = 0.83$), with moderate correlations observed for *PIK3CA* ($r = 0.45$) and *TP53* ($r = 0.36$). Lower *TP53* protein abundance levels compared to mRNA levels were especially prevalent in luminal tumours, suggesting post-transcriptional regulatory mechanisms such as proteasomal degradation. To explore this hypothesis, a search was made for E3 ligases that showed negative correlation to p53 protein (Supplementary Table 8). These analyses identified *UBE3A* ($r = -0.42$; adjusted P value = 0.05) (Extended Data Fig. 4a), an established *TP53* E3 ligase¹⁷. In comparing copy number alterations (CNAs), RNA, and protein levels for *GATA3*, copy number gains in chromosome 10q were anticorrelated with RNA and protein levels in basal-like tumours. This observation prompted a search for other gains or losses that were anticorrelated with RNA and/or protein levels (see Extended Data Fig. 4b for further analyses). Overall, six genes were identified that significantly anticorrelated at a false discovery rate ($FDR < 0.05$) on both RNA and protein levels to their CNA signals (Extended Data Fig. 4b). *GATA3* amplification on 10q in basal-like breast cancer showed the strongest anticorrelation, followed by the hexosamine and glycolysis pathway enzymes *GFPT2* and *HK3*, which are upregulated in basal-like breast cancer despite being subjected to frequent chromosomal deletion on 5q. Global analysis of the correlation of mRNA-to-protein yielded a median Pearson value of $r = 0.39$, with 6,135 out of 9,302 mRNA–protein pairs (66.0%) correlating significantly at an $FDR < 0.05$ (Extended Data Fig. 4c, Supplementary Table 9 and Supplementary Discussion). Similar to a previous colon cancer analysis⁶, metabolic functions such as amino acid, sugar and fatty acid metabolism were found to be enriched among positively correlated genes¹⁸ whereas ribosomal, RNA polymerase and mRNA splicing functions were negatively correlated. Overall these analyses demonstrate the utility of global proteome correlation analysis for both confirmation of suspected regulatory mechanisms and identification of candidate regulators meriting further investigation.

Copy number alterations

To determine the consequences of CNAs on mRNA, protein, and phosphoprotein abundance, both in '*cis*' on genes within the aberrant locus and in '*trans*' on genes encoded elsewhere, univariate correlation analysis was used as previously described⁶. A total of 7,776 genes with CNA, mRNA and protein measurements were analysed by calculating Pearson correlation and associated statistical significance (Benjamini–Hochberg-corrected P value) for all possible CNA–mRNA and CNA–protein pairs (Fig. 2a, Supplementary Table 10, Extended Data Fig. 5a, see Methods). For the phosphoproteome, 4,472 CNA–phosphoprotein pairs were analysed (Extended Data Fig. 5b). Significant positive correlations (*cis*) were observed for 64% of all CNA–mRNA, 31% of all CNA–protein, and 20% of all CNA–phosphoprotein pairs Fig. 2b. Proteins and phosphoproteins correlated in *cis* to CNAs were, for the most part, a subset of the *cis*-effects observed in mRNA–CNA correlation (Fig. 2b, Supplementary Table 10). The fractional difference of well-annotated oncogenes and tumour suppressor genes among the significantly *cis*-correlated CNA–mRNA and CNA–protein gene pairs was analysed. On the basis of a reference list of 487 oncogenes and tumour suppressors (Supplementary Table 10), these cancer-relevant genes occur 37.6% more frequently in the subset of genes that correlate both on CNA–mRNA and CNA–protein levels than in the subset that only correlate on CNA–mRNA but not on CNA–protein levels (Fisher exact P value = 0.02). This suggests that CNA events with a tumour-promoting outcome more likely lead to *cis*-regulatory effects on both the protein and mRNA level, whereas CNA events with no documented role in tumorigenesis are more likely to be neutralized on the protein level than on the RNA level. *Trans*-effects (Fig. 2a) appear as vertical bands, with accompanying frequency histograms (in blue) highlighting 'hot spots' of significant *trans*-effects. Using a minimum threshold of 50 *trans*-affected genes, 68% of the tested genes were associated with *trans*-effects on the mRNA level, whereas only 13% were associated with effects on the protein level and 8% on the phosphoprotein level. Importantly, CNA–protein correlations appeared to be a reduced representation of CNA–mRNA correlations. Furthermore, for many CNA regions, correlations were more directionally uniform on the protein level than on the mRNA level. CNA regions exhibiting the most *trans*-associations at the protein level were found on chromosomes 5q (loss of heterozygosity (LOH) in basal; gain in luminal B), 10p (gain in basal), 12 (gain in basal), 16q (luminal A deletion), 17q (luminal B amplification), and 22q (LOH in luminal and basal) (Extended Data Fig. 5a).

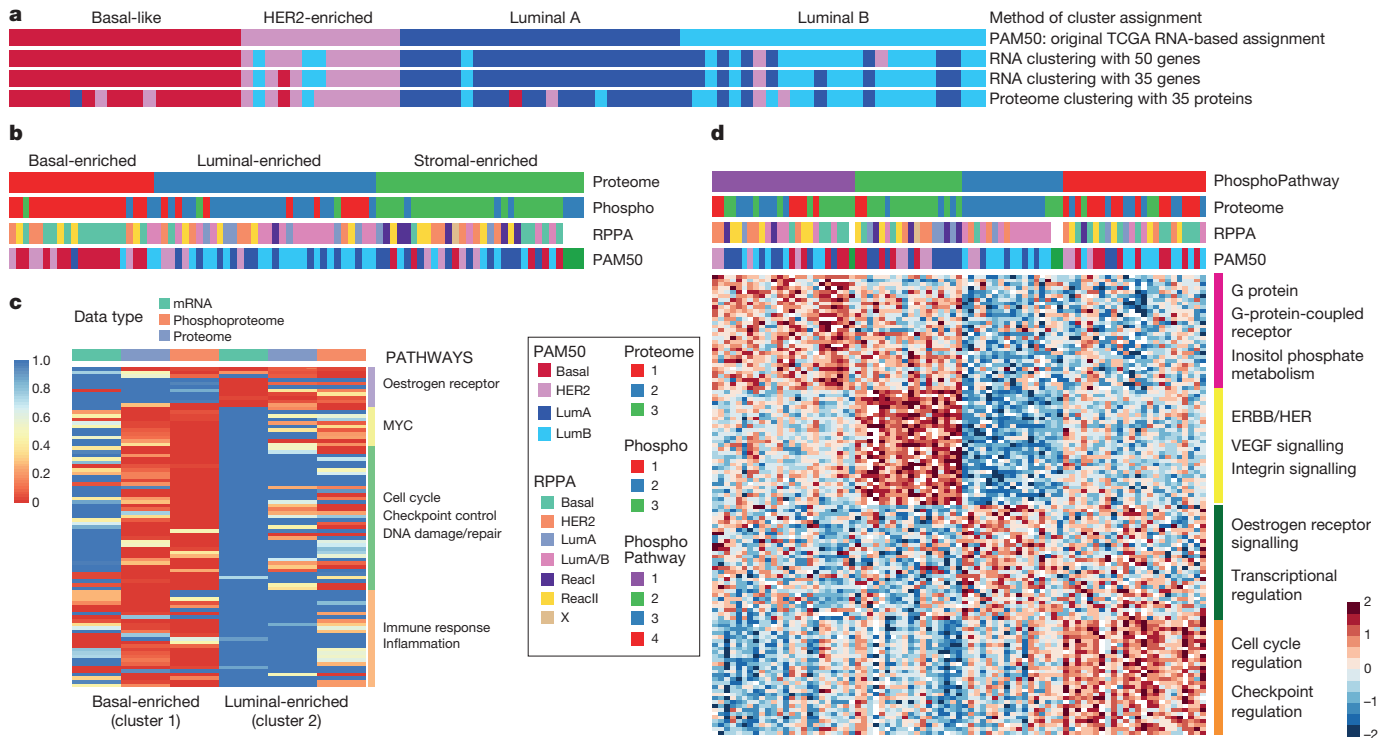


Figure 3 | Proteomic and phosphoproteomic subtypes of breast cancer and subtype-specific pathway enrichment. **a**, Unsupervised clustering of RNA-seq and proteomics data restricted to PAM50 genes and subset of 35 detected proteins reveal high similarity to PAM50 (TCGA) sample annotation. **b**, K-means consensus clustering of proteome and phosphoproteome data identifies basal-enriched, luminal-enriched,

Trans-associations are not necessarily direct consequences of the chromosomal aberration. For example, as 5q loss occurs in at least 50% of basal-like breast cancers¹⁹, many of the *trans*-effects involve genes that mark the basal subtype. To identify candidate driver genes with copy number alterations that are direct drivers of *trans*-effects, results were compared with functional knockdown data on 3,797 genes in the Library of Integrated Network-based Cellular Signatures (LINCS) database (<http://www.lincsproject.org/>)^{20–22}. For any given gene with copy number alterations ('CNA-gene'), sets of genes were identified corresponding to proteins that changed where there was gain ('CNA-gain *trans*-gene set') or loss ('CNA-loss *trans*-gene set'). These gene sets were then compared to the effects of gene knockdown in the LINCS database (see Supplementary Methods). Queries for 502 different CNA genes meeting the criteria defined above identified 10 CNA genes that could be functionally connected to both CNA-gain and CNA-loss *trans*-protein-level effects (Extended Data Fig. 5c, Supplementary Table 11). A permutation-based approach implemented to test significance (see Supplementary Methods) yielded an FDR < 0.05 for 10 genes affected by both CNA gains and losses (Fig. 2c). These proteins were defined as potential regulatory candidates for the CNA *trans*-effects observed on the proteome level in this study, as in a gene-dependent manner an average of 17% of these *trans*-effects were consistent with the knockdown profiles. Notably, the established oncogenic receptor tyrosine kinase *ERBB2* was functionally connected only to CNA gain *trans*-effects (Supplementary Table 11). The E3 ligase *SKP1* (ref. 23) and the ribonucleoprotein export factor *CETN3*, both located on chromosome arm 5q with frequent losses in basal-like breast cancer and less frequent gains in luminal B breast cancer, were detected as potential regulators affecting the expression of the tyrosine kinase and therapeutic target EGFR, and *SKP1* also was linked to SRC (Extended Data Fig. 5d). Another potential regulator, *FBXO7* (a substrate

and stromal-enriched subgroups. **c**, GSEA highlights sets of pathways significantly differential between basal-enriched and luminal-enriched tumours (detailed in Extended Data Fig. 7b). ReactI and ReactII, reactive type I and II, respectively. **d**, K-means consensus clustering performed on pathways derived from single sample GSEA analysis of phosphopeptide data identifies four distinct clusters.

recognition component of the SCF (SKP1-CUL1-F-box protein)-type E3 ubiquitin ligase complex), was affected mostly by LOH events on chromosome 22q. Interestingly, in a recent human interaction proteome study, SKP1 and FBXO7 were listed as interaction partners²⁴.

Clustering and network analyses

Transcriptional profiling has converged on four major breast cancer subtypes: luminal A, luminal B, basal and HER2-enriched^{1,9}. To investigate the extent to which the PAM50 'intrinsic' breast cancer classification scheme is reflected or refined on the proteome level in the CPTAC samples, clustering analyses were first restricted to the reduced set of PAM50 genes. When RNA data for the 50 PAM50 genes were clustered directly (without using a classifier), the clustering was similar to the TCGA PAM50 annotation (second annotation bar in Fig. 3a). Restricting both the RNA and proteome data to the set of 35 PAM50 genes observed in the proteome produced a similar result (bottom two annotation bars in Fig. 3a), and all of the major PAM50 groups were recapitulated in the proteome almost as well as in the RNA data. This indicates that although different tissue sections of the same tumours were used for RNA-seq and protein analysis, very similar subtype-defining features can be observed in both data types. Global proteome and phosphoproteome data were then used to identify proteome subtypes in an unsupervised manner. Consensus clustering identified basal-enriched, luminal-enriched, and stromal-enriched clusters (Extended Data Figs 6a–d, 7a). Unlike the clustering observed with PAM50 genes, mRNA-defined HER2-enriched tumours were distributed across these three proteomic subgroups. The basal-enriched and luminal-enriched groups showed a strong overlap with the mRNA-based PAM50 basal-like and luminal subgroups, whereas stromal-enriched proteome subtype represented a mix of all PAM50 mRNA-based subtypes, and has a significantly

enriched stromal signature (Extended Data Fig. 3e). Among the stromal-enriched tumours there was strong representation of reactive type I tumours, as classified by RPPA (Supplementary Table 12), showing agreement between the RPPA and mass-spectrometry-based protein analyses for the detection of a tumour subgroup characterized by stromal gene expression¹.

As the basal- and luminal-enriched proteome subgroups are coherent, pathway analyses were conducted on these two subtypes, using the stromal-enriched subgroup as a control to assess specificity (Fig. 3c, Extended Data Fig. 7b, Supplementary Table 13). The luminal-enriched subgroup was exclusively enriched for oestradiol- and *ESR1*-driven gene sets. In contrast, multiple gene sets were enriched and upregulated specifically in the basal-like tumours. Particularly extensive basal-like enrichment was seen for *MYC* target genes; for cell cycle, checkpoint, and DNA repair pathways including regulators *AURKA/B*, *ATM*, *ATR*, *CHEK1/2*, and *BRCA1/2*; and for immune response/inflammation, including T-cell, B-cell, and neutrophil signatures. The complementarity of transcriptional, proteomic, and phosphoproteomic data was also highlighted in these analyses (Extended Data Fig. 7c, d).

Using phosphorylation status as a proxy for activity, phosphoproteome profiling can theoretically be used to develop a signalling-pathway-based cancer classification. *K*-means consensus clustering was therefore performed on pathways derived from single sample gene set enrichment analysis (GSEA) of phosphopeptide data (Methods, Supplementary Tables 14 and 15). Of four robustly segregated groups, subgroups 2 and 3 substantially recapitulated the stromal- and luminal-enriched proteomic subgroups, respectively (Fig. 3d, Extended Data Fig. 8a). Subgroup 4 included a majority of tumours from the basal-enriched proteomic subgroup, but was admixed particularly with luminal-enriched samples. This subgroup was defined by high levels of cell cycle and checkpoint activity. All basal and a majority of non-basal samples in this subgroup had *TP53* mutations. Consistent with high levels of cell cycle activity, a multivariate kinase-phosphosite abundance regression analysis highlighted CDK1 as one of the most highly connected kinases in this study (Extended Data Fig. 8b, Supplementary Table 16). Subgroup 1 was a novel subgroup defined exclusively in the phosphoproteome pathway activity domain, with no enrichment for either proteomic or PAM50 subtypes. It was defined by G protein, G-protein-coupled receptor, and inositol phosphate metabolism signatures, as well as ionotropic glutamate signalling (Fig. 3d). Co-expression patterns among genes/proteins across different subgroups were also analysed using a Joint Random Forest method²⁵ that identified network modules, such as an MMP9 module, with different interaction patterns between basal-enriched and luminal-enriched subgroups. These latter patterns appeared specific to the proteome-level data (Extended Data Fig. 8c–f, Supplementary Table 17 and Supplementary Methods).

Phosphosite markers in *PIK3CA*- and *TP53*-mutated tumours

TP53 and *PIK3CA* are the most recurrently mutated genes in breast cancer, with frequencies for *PIK3CA* at 43% in luminal tumours and for *TP53* at 84% in basal-like tumours¹. Most of the *PIK3CA* missense mutations were gain of function mutations and therefore were expected to lead to activation of the PI3K signalling cascade, but the extent to which this occurs has been controversial and it is unclear which pathway components are effectors^{26,27}. Marker selection analysis was therefore performed for upregulated phosphosites in *PIK3CA*-mutated tumours. In total, 62 phosphosites were identified that were positively associated with *PIK3CA* mutation (FDR < 0.05), including the kinases RPS6KA5 and EIF2AK4 (Extended Data Fig. 9a, Supplementary Table 18). Calculating the average phosphorylation signal of these marker phosphosites provided a read-out for PI3K pathway activity in *PIK3CA*-mutated tumours, with 15 of the 26 mutated tumours (58%) exhibiting an activated *PIK3CA* mutation signature. Of note, the identified

PIK3CA mutant phosphoproteome signature was activated in all tumours harbouring helical domain *PIK3CA* mutations, but only 2 of 10 tumours harbouring kinase domain mutations. To test if the identified differences in the phosphoproteome of PI3K mutant versus wild-type tumours could be explained by mutation of *PIK3CA*, the tumour data were compared to phosphosite signatures derived from isogenic *PIK3CA* mutant cell lines²⁸ (Extended Data Fig. 9b, Supplementary Table 18). There was an enrichment of signatures derived from helical-domain-mutated isogenic cell lines, but not from kinase-domain-mutated cells, supporting the observations in primary tumours.

The same strategy was used to identify phosphorylation signalling events connected to *TP53* mutation. A total of 56 phosphosites upregulated in *TP53*-mutated tumours were identified that were independent of basal-like subtype association (Extended Data Fig. 9c, Supplementary Table 18). Using the average phosphorylation signal of these marker phosphosites as a proxy for *TP53*-mutation-driven cell cycle control, 22 of 41 mutated tumours (54%) showed upregulated signals. This *TP53* mutant phosphosignature was somewhat enhanced in tumours in which mutations occurred almost exclusively in the DNA-binding region compared to those with nonsense/frameshift mutations. In addition to the well-described checkpoint kinase *CHEK2*, significantly upregulated phosphosites were identified for the kinases *MASTL* and *EEF2K* in *TP53*-mutated tumours. Single-sample GSEA analysis of isogenic p53-mutant phosphosignatures showed an enrichment of a phosphosignature derived from *R273H*-mutated isogenic cells (Extended Data Fig. 9d), confirming the pronounced effect of missense mutations in the DNA-binding region on phosphorylation pathways.

Kinase gene amplification and subtype-specific activation

CNAs span many driver gene candidates and RNA expression has been frequently used to narrow candidate nominations. Proteogenomic analysis should further promote this nomination process. In candidate refinement, a focus on protein kinases is warranted, as many are drug targets. An in-depth proteogenomic pipeline was developed that flagged kinases, expression levels of which were at least 1.5 interquartile ranges higher than the median (Supplementary Table 19). A proteogenomic circos-like²⁹ plot (termed a 'pircos' plot) was used to map these outlier values onto the genome (Fig. 4a, b, Extended Data Fig. 10a). The *ERBB2* locus showed the strongest effect of increased phosphoprotein levels associated with gene-amplification-driven RNA and protein over expression (Fig. 4a). The kinase CDK12 is a positive transcriptional regulator of homologous recombination repair genes with its partner cyclin K³⁰, and is often encompassed by the *ERBB2* amplicon. This gene was also found to be upregulated at the RNA, protein, and phosphosite level indicating that *CDK12* is highly active in the majority of *ERBB2*-positive tumours (Fig. 4a). The analysis of the *ERBB2* amplicon also uncovered co-outlier phosphorylation status for *MED1*, *GRB7*, *MSL1*, *CASC3* and *TOP2A*, all previously described in association with *ERBB2* amplification. To better understand the downstream effects of *ERBB2* amplification, additional phosphosite outliers were identified in 41 known *ERBB2* signalling genes for the 15 samples that had *ERBB2* phosphosite outlier expression (Extended Data Fig. 10b).

These canonical findings stimulated a proteogenomic analysis to identify additional outlier kinases in the breast cancer genome. A proteogenomic dissection of chromosome 11q based on *PAK1* amplification (Fig. 4b, c), a breast cancer driver kinase³¹, illustrated that *PAK1* is hyperphosphorylated in *PAK1*-amplified tumours, along with *CLNS1A*, *RFS1* and *GAB2* (ref. 32). Additional examples of outlier kinases included *PTK2* and *RIPK2* in association with amplification of chromosome 8q (Fig. 4c, Extended Data Fig. 10a, c). *PAK1* and *TLK2* (17q23) appear to be luminal-breast-cancer-specific events (Fig. 4c, Extended Data Fig. 10c). To further examine whether outlier kinases were breast cancer subtype-specific independent of amplification

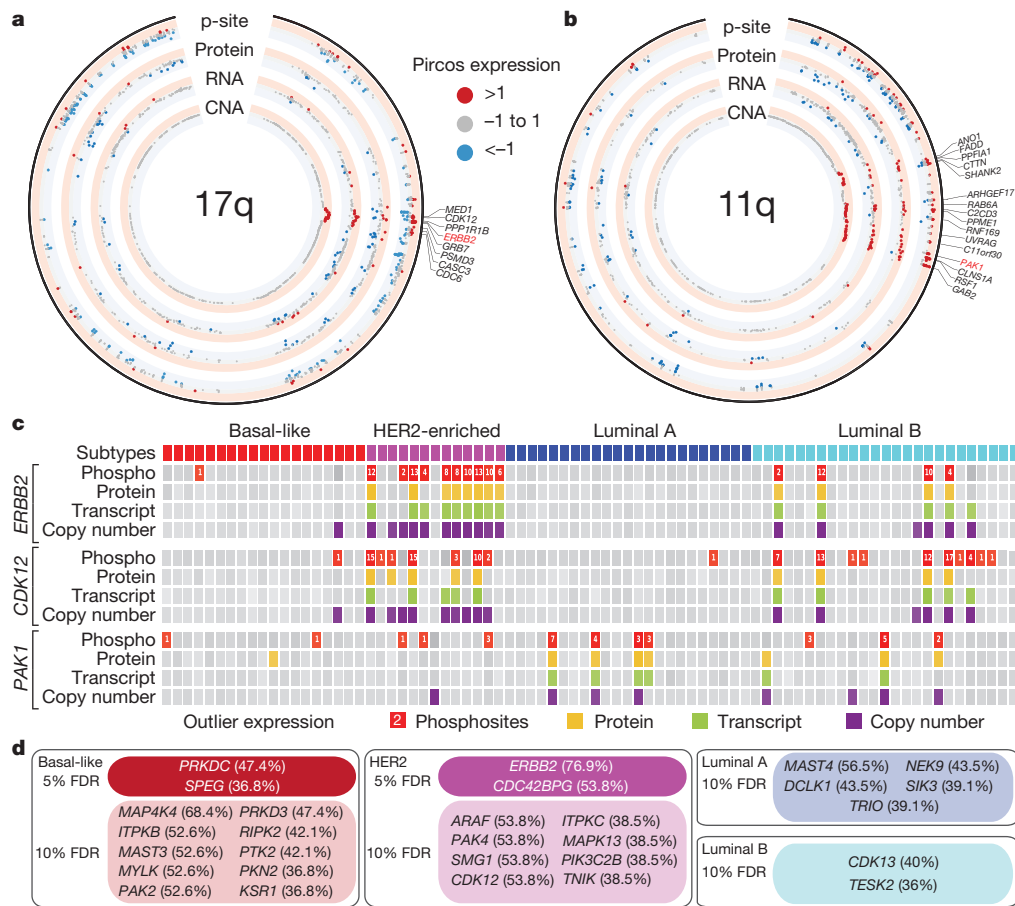


Figure 4 | Example analyses of aberrantly regulated kinases in human breast cancer. **a, b**, pircos (proteogenomics circos) plots showing CNA, RNA, protein, and phosphosite expression for 17 tumours with amplification in 17q (*ERBB2* CNA >1) and 8 tumours with amplification in 11q (*PAK1* CNA >1). Labelled genes have CNA >1 and phosphosite >1. **c**, Proteogenomic outlier expression analysis for *ERBB2*, *CDK12*, and

PAK1. Samples with outlier phosphosite (red), protein (yellow), RNA (green) and copy number (purple) expression are shown. Phosphosite squares indicate per-sample outlier phosphosites. **d**, Outlier kinase events by PAM50 subtype (>35% of subtype samples contain a phosphosite outlier; <10% FDR using Benjamini–Hochberg-adjusted *P* values).

status, the Benjamini–Hochberg-corrected probability was calculated of finding the number of phosphosite outliers within a subtype, given the total number of outliers across all subtypes, the subtype sample size and the total sample size (Fig. 4d). These analyses led to the expected identification of *ERBB2* in the HER2-enriched subtype at the 5% FDR level, as well as the new finding of *CDC42BPG* (MRGK γ), an effector kinase for RHO-family GTPases³³. In basal-like breast cancer, two kinases, *PRKDC* and *SPEG*, were significant at the 5% FDR level. *PRKDC* is a non-homologous end-joining factor that can be phosphorylated by ATM kinase, and is therefore a logical finding in this disease subset³⁴. However *SPEG*, a kinase associated with severe dilated cardiomyopathy when suppressed³⁵, has not been previously reported in association with breast cancer. A larger number of subtype-specific kinases were detected at the 10% FDR level, several of which have recently described relevance in breast cancer, including *PRKD3* in basal-like breast cancer³⁶, the LKB-regulated *SIK3* in luminal A breast cancer³⁷ and *CDK13* in luminal B breast cancer, which, similar to *CDK12*, can interact with cyclin K³⁰.

Discussion

The breadth and depth of proteomic and phosphoproteomic analyses displayed in this study demonstrates the strength of mass-spectrometry-based proteomics, but also some of the limitations inherent in proteolytic peptide sequencing (see Supplementary Discussion). An example of how high-dimensional proteomic analysis provides insight into unresolved genomic issues concerns the study of loss of the long arm of chromosome 5 (5q). Analysis of RNA and

protein correlations narrowed the list of potential *trans*-deregulated proteins. Orthogonal candidate screening using functional genomics methodologies identified loss of *CETN3* and *SKP1* as potential *trans*-regulators, with upregulation of EGFR as a downstream consequence in basal-like breast cancers. Although further experimental evidence must be sought for these proposed regulatory relationships, the SKP1–Cullin complex has already been linked to EGFR activation in glioma³⁸. Unfortunately, EGFR targeting has not proven to be effective therapy in basal-like breast cancer to date³⁹. This might be due to the fact the *SKP1* loss deregulates multiple targets, therefore mandating a much broader inhibitory strategy.

It is recognized that *PIK3CA* mutations do not strongly activate canonical downstream effectors²⁸. Mass-spectrometry-based phosphoproteomics provides an opportunity for unbiased examination of downstream signalling events dependent on *PIK3CA* mutational activation. These studies revealed that common *PIK3CA* mutations affect a large number of targets with diverse functionalities including the kinases RPS6KA5 and EIF2AK4. Thus, the data and analyses reported here extend our knowledge of the effectors that promote tumorigenesis in response to constitutive activation of PI3 kinase. Similarly, *TP53*-mutation-associated phosphopeptides point towards novel functionalities, including regulation of the kinases MASTL and EEF2K.

A central goal in breast cancer research has been the identification of druggable kinases beyond HER2. Candidate genes that exhibited similar gene-amplification-driven proteogenomic patterns to *ERBB2* included *CDK12*, *TLK2*, *PAK1* and *RIPK2*. The proteogenomic link with gene amplification was particularly strong for *CDK12*, in keeping

with its location in the *ERBB2* amplicon, whereas the strengths of correlation between DNA amplification, RNA, protein, and phosphoprotein for the other examples were more variable. The presence of activated *CDK12* in the *ERBB2* amplicon might explain why tumours arising in *BRCA1* carriers are usually *ERBB2*-negative. As a positive transcriptional regulator of *BRCA1* and multiple FANC family members, *CDK12* promotes DNA repair by homologous recombination. *CDK12* amplification would, therefore, oppose the functional effects of *BRCA1* haploinsufficiency during tumour evolution³⁰. Overall, multiple outlier kinases generate testable therapeutic hypotheses for which enabling inhibitors are in development. For example, PAK1 has recently been confirmed to be a therapeutic target and poor prognosis factor in luminal breast cancer⁴⁰.

Although incomplete outcome data and the remarkable heterogeneity of breast cancer are further relevant constraints, the number of TCGA specimens analysed here is insufficient to support conclusive clinical correlations. Only 8 deaths occurred among the 77 patients, which are too few to provide sufficient statistical power for association analysis. Adequately powered MS/MS-based clinical investigation will require microscaled discovery or targeted approaches⁴¹, especially given the highly limited amount of patient material available from clinical trials and the mostly formalin-fixed nature of the specimens. The current analysis is therefore centred on biological findings and correlations, with orthogonal validation and false discovery concerns addressed through an examination of cell-line databases of the effects of individual gene perturbations. Typical of a multi-tiered analysis of this complexity, there are many hypotheses to test, and many findings that require further investigation.

In conclusion, this study provides a high-quality proteomic resource for human breast cancer investigation, and illustrates technologies and analytical approaches that provide an important new opportunity to connect the genome to the proteome. Larger-scale exploration of discovery proteomics in the clinical setting will require improvements in clinical investigation, including acquisition of adequate amounts of optimally collected tumour tissue both before and during therapy as well as advances in MS/MS proteomics to reduce sample input and increase sensitivity for low abundance proteins and modified peptides.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 2 July 2015; accepted 13 April 2016.

Published online 25 May 2016.

1. Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. *Nature* **490**, 61–70 (2012).
2. Curtis, C. *et al.* The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature* **486**, 346–352 (2012).
3. van 't Veer, L. J. *et al.* Gene expression profiling predicts clinical outcome of breast cancer. *Nature* **415**, 530–536 (2002).
4. Chin, K. *et al.* Genomic and transcriptional aberrations linked to breast cancer pathophysiologies. *Cancer Cell* **10**, 529–541 (2006).
5. Ellis, M. J. *et al.* Connecting genomic alterations to cancer biology with proteomics: the NCI Clinical Proteomic Tumor Analysis Consortium. *Cancer Discov.* **3**, 1108–1112 (2013).
6. Zhang, B. *et al.* Proteogenomic characterization of human colon and rectal cancer. *Nature* **513**, 382–387 (2014).
7. Perou, C. M. *et al.* Molecular portraits of human breast tumours. *Nature* **406**, 747–752 (2000).
8. Sorlie, T. *et al.* Repeated observation of breast tumor subtypes in independent gene expression data sets. *Proc. Natl Acad. Sci. USA* **100**, 8418–8423 (2003).
9. Parker, J. S. *et al.* Supervised risk predictor of breast cancer based on intrinsic subtypes. *J. Clin. Oncol.* **27**, 1160–1167 (2009).
10. Li, S. *et al.* Endocrine-therapy-resistant *ESR1* variants revealed by genomic characterization of breast-cancer-derived xenografts. *Cell Reports* **4**, 1116–1130 (2013).
11. Polyak, K. Heterogeneity in breast cancer. *J. Clin. Invest.* **121**, 3786–3788 (2011).
12. Bertos, N. R. & Park, M. Breast cancer — one term, many entities? *J. Clin. Invest.* **121**, 3789–3796 (2011).
13. Symmans, W. F., Liu, J., Knowles, D. M. & Inghirami, G. Breast cancer heterogeneity: evaluation of clonality in primary and metastatic lesions. *Hum. Pathol.* **26**, 210–216 (1995).

14. Yoshihara, K. *et al.* Inferring tumour purity and stromal and immune cell admixture from expression data. *Nat. Commun.* **4**, 2612 (2013).
15. Mertins, P. *et al.* Ischemia in tumors induces early and sustained phosphorylation changes in stress kinase pathways but does not affect global protein levels. *Mol. Cell Proteomics* **13**, 1690–1704 (2014).
16. Ruggles, K. V. *et al.* An analysis of the sensitivity of proteogenomic mapping of somatic mutations and novel splicing events in cancer. *Mol. Cell Proteomics* **15**, 1060–1071 (2015).
17. Scheffner, M., Huibregtse, J. M., Vierstra, R. D. & Howley, P. M. The HPV-16 E6 and E6-AP complex functions as a ubiquitin-protein ligase in the ubiquitination of p53. *Cell* **75**, 495–505 (1993).
18. Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. USA* **102**, 15545–15550 (2005).
19. Silva, G. O. *et al.* Cross-species DNA copy number analyses identifies multiple 1q21-q23 subtype-specific driver genes for breast cancer. *Breast Cancer Res. Treat.* **152**, 347–356 (2015).
20. Lamb, J. *et al.* The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. *Science* **313**, 1929–1935 (2006).
21. Peck, D. *et al.* A method for high-throughput gene expression signature analysis. *Genome Biol.* **7**, R61 (2006).
22. Duan, Q. *et al.* LINCS Canvas Browser: interactive web app to query, browse and interrogate LINCS L1000 gene expression signatures. *Nucleic Acids Res.* **42**, W449–W660 (2014).
23. Nakayama, K. I. & Nakayama, K. Ubiquitin ligases: cell-cycle control and cancer. *Nat. Rev. Cancer* **6**, 369–381 (2006).
24. Hein, M. Y. *et al.* A human interactome in three quantitative dimensions organized by stoichiometries and abundances. *Cell* **163**, 712–723 (2015).
25. Petralia, F., Song, W. M., Tu, Z. & Wang, P. New method for joint network analysis reveals common and different coexpression patterns among genes and proteins in breast cancer. *J. Proteome Res.* **15**, 743–754 (2016).
26. Loi, S. *et al.* *PIK3CA* mutations associated with gene signature of low mTORC1 signaling and better outcomes in estrogen receptor-positive breast cancer. *Proc. Natl Acad. Sci. USA* **107**, 10208–10213 (2010).
27. Vasudevan, K. M. *et al.* AKT-independent signaling downstream of oncogenic *PIK3CA* mutations in human cancer. *Cancer Cell* **16**, 21–32 (2009).
28. Wu, X. *et al.* Activation of diverse signalling pathways by oncogenic *PIK3CA* mutations. *Nat. Commun.* **5**, 4961 (2014).
29. Krzywinski, M. *et al.* Circos: an information aesthetic for comparative genomics. *Genome Res.* **19**, 1639–1645 (2009).
30. Blazek, D. *et al.* The Cyclin K/Cdk12 complex maintains genomic stability via regulation of expression of DNA damage response genes. *Genes Dev.* **25**, 2158–2172 (2011).
31. Shrestha, Y. *et al.* *PAK1* is a breast cancer oncogene that coordinately activates MAPK and MET signaling. *Oncogene* **31**, 3397–3408 (2012).
32. Chen, Y. *et al.* Identification of druggable cancer driver genes amplified across TCGA datasets. *PLoS One* **9**, e98293 (2014).
33. Prudnikova, T. Y., Rawat, S. J. & Chernoff, J. Molecular pathways: targeting the kinase effectors of RHO-family GTPases. *Clin. Cancer Res.* **21**, 24–29 (2015).
34. Jiang, W. *et al.* Differential phosphorylation of DNA-PKcs regulates the interplay between end-processing and end-ligation during nonhomologous end-joining. *Mol. Cell* **58**, 172–185 (2015).
35. Agrawal, P. B. *et al.* SPEG interacts with myotubularin, and its deficiency causes centronuclear myopathy with dilated cardiomyopathy. *Am. J. Hum. Genet.* **95**, 218–226 (2014).
36. Borges, S. *et al.* Effective Targeting of estrogen receptor-negative breast cancers with the protein kinase D inhibitor CRT0066101. *Mol. Cancer Ther.* **14**, 1306–1316 (2015).
37. Walkinshaw, D. R. *et al.* The tumor suppressor kinase LKB1 activates the downstream kinases SIK2 and SIK3 to stimulate nuclear export of class IIa histone deacetylases. *J. Biol. Chem.* **288**, 9345–9362 (2013).
38. Jiang, X. *et al.* Numb regulates glioma stem cell fate and growth by altering epidermal growth factor receptor and Skp1-Cullin-F-box ubiquitin ligase activity. *Stem Cells* **30**, 1313–1326 (2012).
39. Carey, L. A. *et al.* TBCRC 001: randomized phase II study of cetuximab in combination with carboplatin in stage IV triple-negative breast cancer. *J. Clin. Oncol.* **30**, 2615–2623 (2012).
40. Ong, C. C. *et al.* Small molecule inhibition of group I p21-activated kinases in breast cancer induces apoptosis and potentiates the activity of microtubule stabilizing agents. *Breast Cancer Res.* **17**, 59 (2015).
41. Carr, S. A. *et al.* Targeted peptide measurements in biology and medicine: best practices for mass spectrometry-based assay development using a fit-for-purpose approach. *Mol. Cell Proteomics* **13**, 907–917 (2014).

Supplementary Information is available in the online version of the paper.

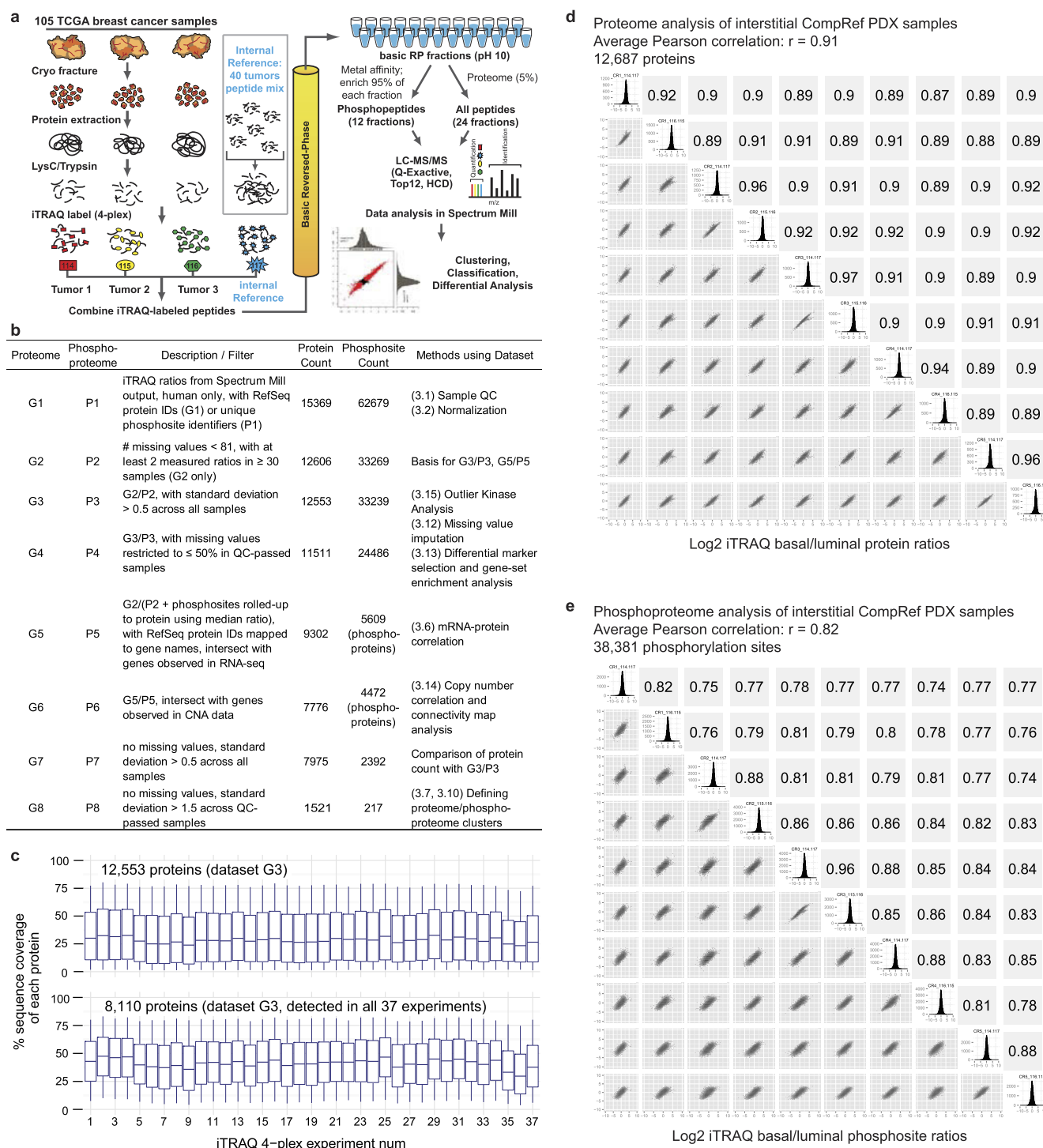
Acknowledgements This work was supported by National Cancer Institute (NCI) CPTAC awards U24CA160034 (Broad Institute; Fred Hutchinson Cancer Research Center), U24CA160036 (Johns Hopkins University), U24CA160019 (Pacific Northwest National Laboratory), U24CA159988 (Vanderbilt University), U24CA160035 (Washington University, St. Louis; University of North Carolina, Chapel Hill). P.W. and F.P. were also supported by SUB-R01GM108711 and MJE by CPRIT grant RR140033. M.J.E. is also a McNair Foundation Scholar. D.F. was supported by Leidos contract 13XS068. Primary genomics data for this study were generated by The

Cancer Genome Atlas pilot project established by the NCI and the National Human Genome Research Institute. Resequencing of select samples conducted in this study was supported by National Cancer Institute (NCI) CPTAC award U24CA160035. Information about TCGA and the investigators and institutions that constitute the TCGA research network can be found at <http://cancergenome.nih.gov/>. We also acknowledge the expert assistance of J. Snider, P. Erdmann-Gilmore and R. Connors for the preparation of the tumour tissues for solubilization. We thank the Alvin J. Siteman Cancer Center at Washington University School of Medicine and Barnes-Jewish Hospital in St. Louis, for the use of the Tissue Procurement Core, which provided accessioning, histologic processing and review for the TCGA samples included in this study. The Siteman Cancer Center is supported in part by an NCI Cancer Center Support Grant #P30 CA91842 (see more at <http://www.siteman.wustl.edu/ContentPage.aspx?id=243#sthash.mEU0QuXx.dpuf>). We also thank the HAMLET Core at The Washington University in St. Louis for providing breast cancer xenograft tumors. The HAMLET Core was supported in part by grants from NIH/NCRR Washington University-ICTS (UL1 RR024992) and Susan G. Komen for the Cure (KG 090422). F.M. was also supported by The Swedish Research Council (Dnr 2014-323). We also thank A. Subramanian, C. Flynn and J. Asiedu at the Broad Institute for their guidance and assistance in accessing LINCS to run a large number of enrichment queries.

Author Contributions P.M., D.R.M., M.A.G., K.R.C., and S.A.C. designed the proteomic analysis experiments, data analysis workflow, and proteomic-genomic data comparisons. P.M., M.A.G., J.W.Q., and S.A.C. directed and performed proteomic analysis of breast tumour and quality control samples. P.M., D.R.M., K.V.R., K.R.C., P.W., X.W., S.C., E.K., F.P., Z.T., J.T.L., M.L.G., M.W., V.Y., K.H., C.L., M.D.M., P.Y., J.W., B.Z., and D.F. performed proteomic-genomic data

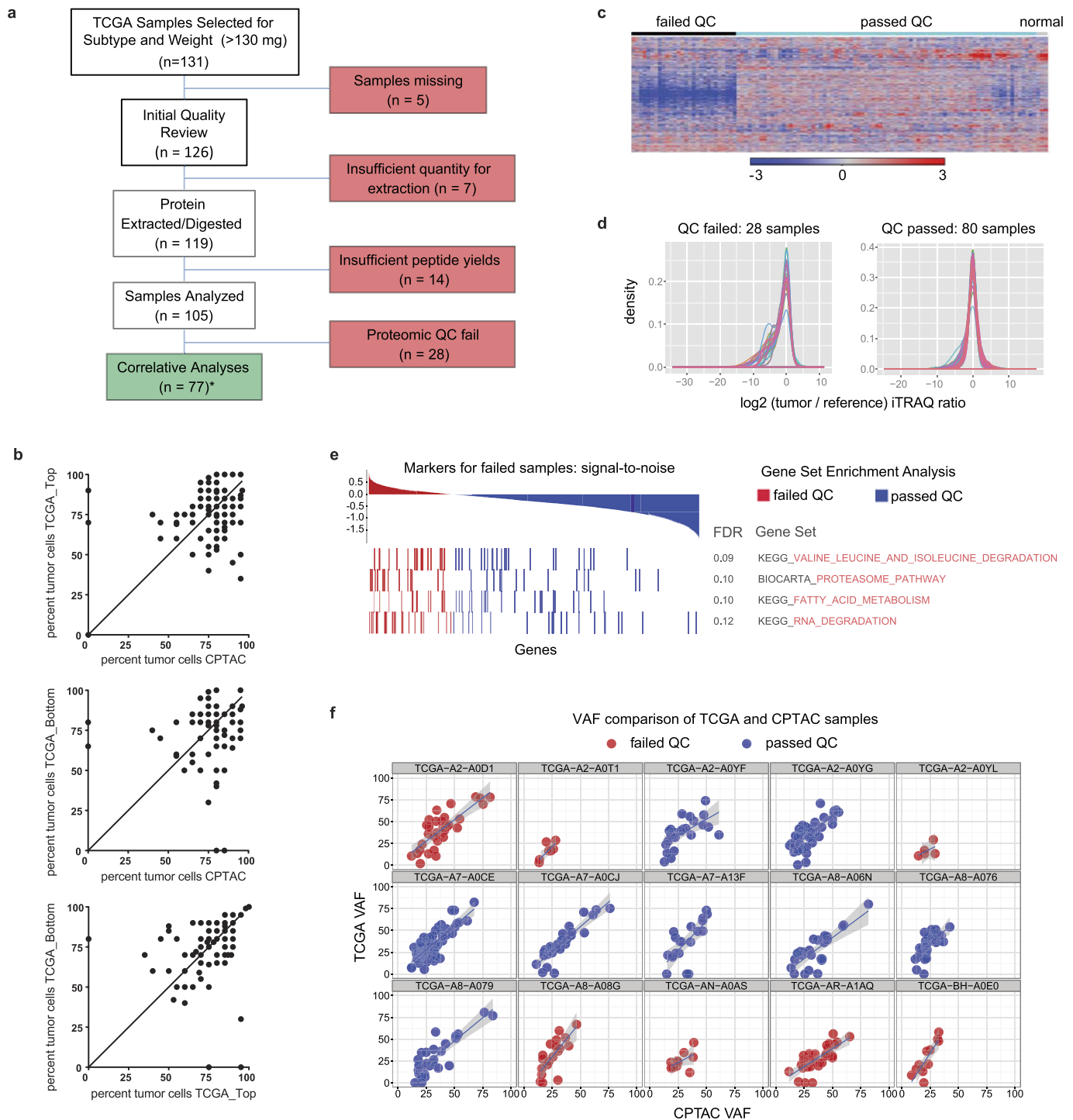
analyses. D.R.M., P.W., and S.J.S. provided statistical support. D.R.M., K.V.R., K.R.C., K.K. and D.F. performed analyses of mass spectrometry data and adapted algorithms and software for data analysis. S.R.D., R.R.T and M.J.E. developed and prepared breast xenografts used as quality control samples. P.M. and F.M. prepared and analyzed cell lines for correlative functional annotation of frequently mutated genes. P.M., D.R.M., M.A.G., and S.A.C. designed strategy for quality control analyses. M.A.G., S.R.D., C.R.K., M.M., and H.R. coordinated acquisition, distribution and quality control evaluation of TCGA tumour samples. P.M., M.A.G., C.M.P., L.D., A.G.P., and M.J.E. interpreted data in the context of breast cancer biology. P.M., D.R.M., M.A.G., K.R.C., P.W., A.G.P., M.J.E. and S.A.C. wrote the manuscript.

Author Information All primary mass spectrometry data are deposited at the CPTAC Data Portal as raw and mzML files and complete protein assembly data sets for public access (<https://cptac-data-portal.georgetown.edu/cptac/s/S029>). In addition, a set of ancillary files such as dataset G1/P1, G3/P3, G4/P4, G5/P5, G7/P7, CNA correlation tables for CNA-mRNA, CNA-proteome and CNA-phosphoproteome, CNA data, and RNA-seq expression data have also been deposited at the CPTAC Data Coordinating Center (DCC). Two browsers for the results: one provides track hubs for viewing the identified peptides in the UCSD genome browser (http://fenyolab.org/cptac_breast_ucsc); the other is an online tool for proteogenomic data exploration, accessed at <http://prot-shiny-vm.broadinstitute.org:3838/BC2016/> (see Supplementary Methods for descriptions). Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to P.M. (pmertins@broadinstitute.org), M.J.E. (Matthew.Ellis@bcm.edu) or S.A.C. (scarr@broad.mit.edu).



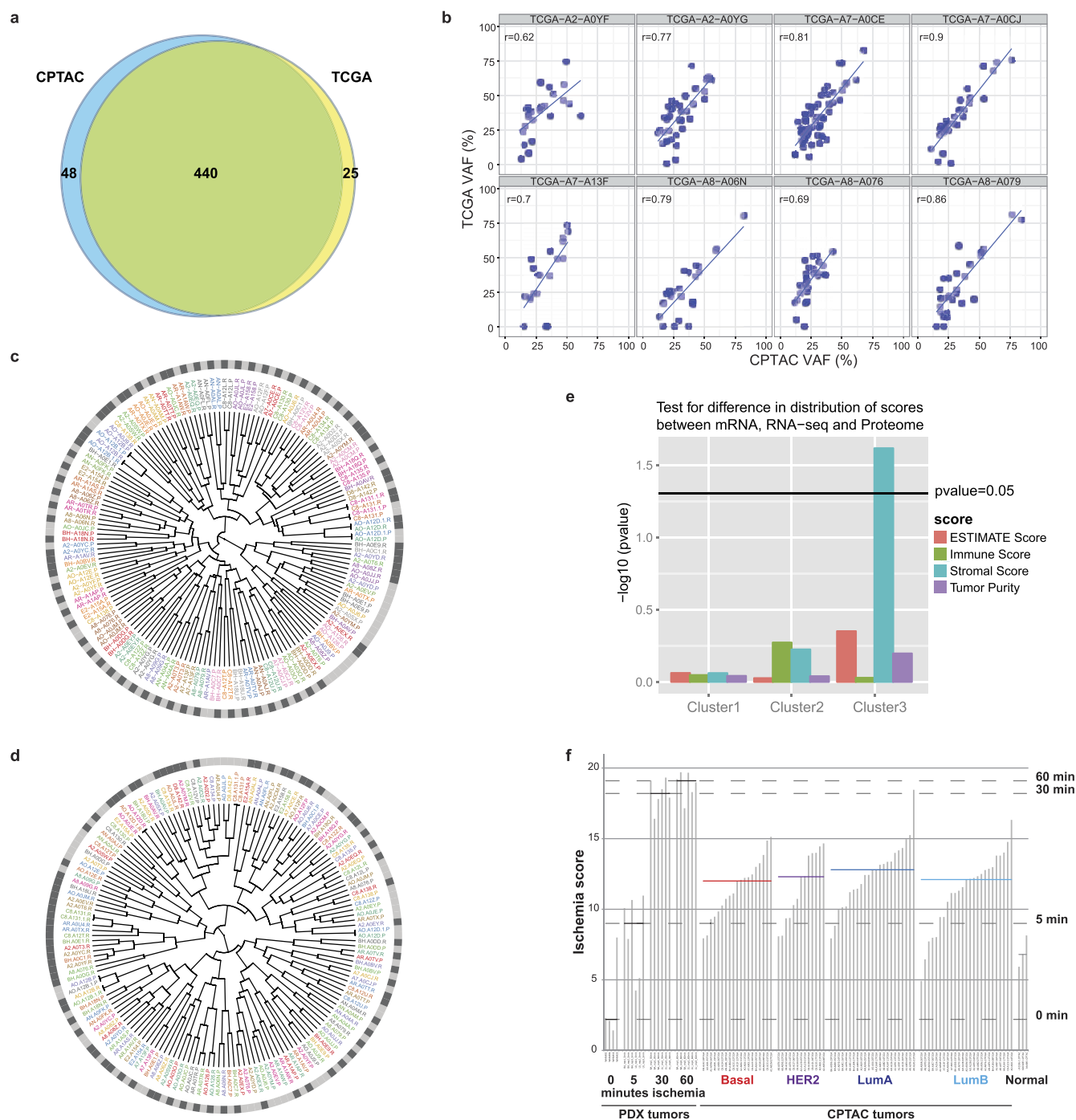
Extended Data Figure 1 | Experimental and data analysis workflows and longitudinal data generation quality control. **a**, iTRAQ 4-plex global proteome and phosphoproteome analysis workflow. 105 TCGA breast tumours were analysed in 35 iTRAQ 4-plex experiments (plus one replicate and one normal sample experiment), with three tumours of different subtypes compared to a fourth common internal reference sample in each experiment. The reference sample comprised 10 individual tumours of each of the four major breast cancer intrinsic subtypes and served as an internal standard for all proteins and phosphoproteins quantified in this study. Each iTRAQ MS/MS spectrum measures a peptide from four samples (3 individual patients and the reference sample mix of 40 patients). More than 400,000 distinct peptides were identified and quantified in ~14 million MS/MS spectra. Personalized tumour-specific protein databases were generated in the QUILTS software package using

whole-exome-sequencing-derived variant calls and RNA-seq-derived transcript information. All mass spectrometry data was analysed using the Spectrum Mill software package. **b**, Overview of proteome and phosphoproteome data sets. The table provides a summary of the data sets used in specific analyses, including the filters applied to derive the proteins and phosphosites/phosphoproteins that constitute each data set; the protein, phosphosite or phosphoprotein count; and the methods that employ the respective data sets. **c**, Distribution of sequence coverage of the identified proteins with tryptic peptides detected by MS/MS, whiskers show the 5–95 percentiles. **d**, **e**, Robust and accurate proteome/phosphoproteome platform. Longitudinal performance was tested by repeated proteome and phosphoproteome analysis of patient-derived xenograft tumours. Scatter plots, histograms and Pearson correlations comparing individual replicate measurements are shown.



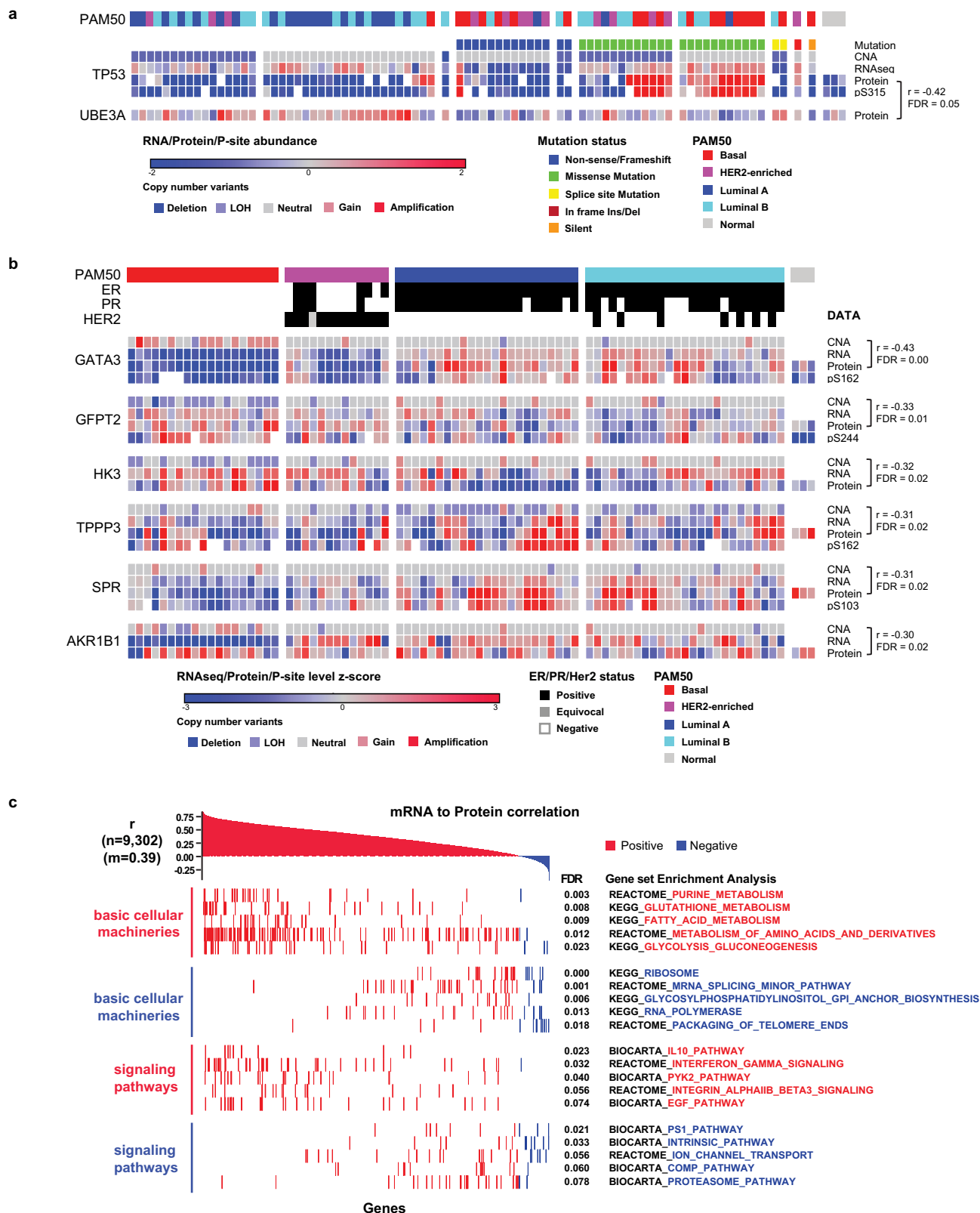
Extended Data Figure 2 | Tumour sample quality control. **a**, Remark diagram showing sample processing and partitioning. Initial quality review encompassed histopathological examination of tissue slices stained with haematoxylin and eosin. *For 3 samples, no tumour cells were seen on histopathology (BH-A0E9, BH-A0C1, A2-A0SW). These samples were nevertheless included in the proteome analysis as other quality control standards were met (see below) and samples with 0% tumour cellularity on top or bottom sections were included in TCGA analyses. **b**, Correlation of TCGA (top or bottom sections) and CPTAC histological assessment of neoplastic cellularity for samples ($n = 105$). The average and range of neoplastic cellularities were identical for CPTAC and TCGA histological assessments. Averages (s.d.) for neoplastic cellularity were 76% (± 17) for CPTAC, 76% (± 15) for TCGA_Top, and 75% (± 18) for TCGA_Bottom histopathology slides (Supplementary Table 2). Note that in three CPTAC cases where no tumour cells were identified by histopathological

assessment, numbers of protein-level somatic variants were similar to all other tumours. The identified mutated proteins were TP53_R273C, NOP58_Q23E, TAGLN2_G154R, TUBA1B_D116H, and MRPL48_I173K (Supplementary Table 5), indicating presence of tumour cells in these samples. **c**, Proteome iTRAQ tumour to internal reference ratio heat map for all CPTAC samples (8,028 proteins without missing values) including passed and failed proteomic quality control (QC) samples. **d**, Global tumour to reference proteome ratio distributions for samples that passed and failed proteomic quality control analysis. **e**, Degradation-related gene sets were enriched in tumours that failed proteomic quality control analysis. **f**, Variant allele frequency (VAF) analysis of re-sequenced CPTAC tumours and comparison to original TCGA data. Overall VAFs for failed quality control samples were lower compared to passed samples suggesting lower purity.



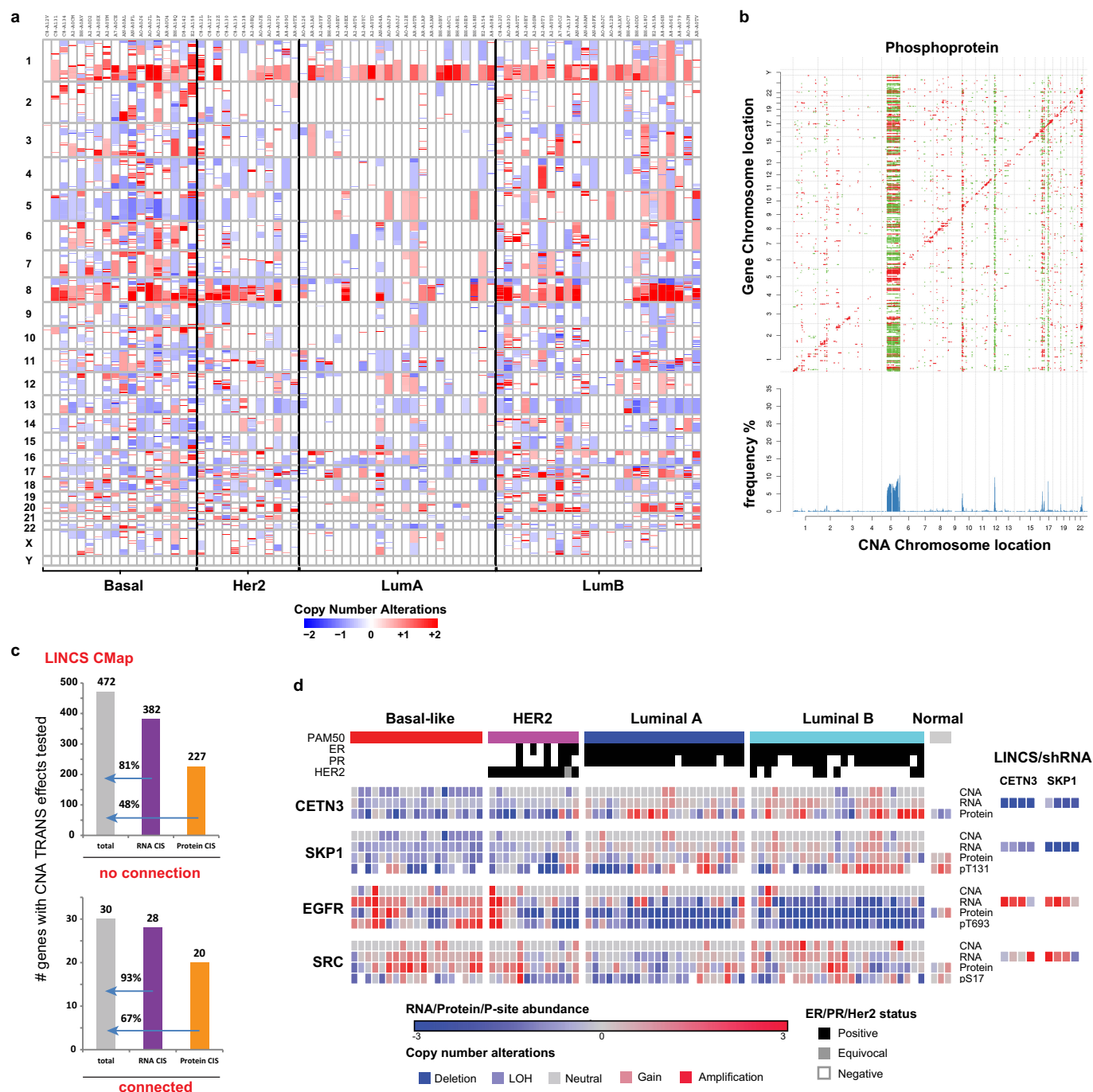
Extended Data Figure 3 | Tumour sample quality control. **a**, There was high concordance (94.6%) between DNA variants reported by TCGA and CPTAC re-sequenced tumours. Most point mutations reported by TCGA could be identified across the eight re-sequenced samples used in the study. **b**, A high overall correlation (mean = 0.77) was observed for the CPTAC VAF (x axis) and TCGA VAF (y axis) across the eight samples used in the study. **c**, Agglomerative hierarchical clustering (Supplementary Methods section 3.8) used to co-cluster protein and RNA tumour expression data after filtering to retain 4,291 proteins and genes with moderate to high protein–RNA correlation (Pearson correlation > 0.4) with results displayed as a circular dendrogram (fanplot). The proteome (.P) and RNA (.R) components of each sample are labelled using the same colour. The outer ring shows proteome samples in light grey and RNA samples in dark grey. High concordance between RNA and protein expression is evident from the colour adjacency in the inner ring and alternating colour in the outer ring showing that RNA and protein components co-cluster for a large proportion of samples (62 out of 80). **d**, Co-clustering of MS/MS and RPPA tumour data. 126 RPPA readouts were mapped to gene names. These

genes were intersected with the genes observed in the MS/MS proteome, filtered to 48 proteins with moderate or higher RPPA–MS/MS protein correlation, and analysed for co-clustering as in **c**. 47 of 80 RPPA–MS/MS protein pairs co-cluster. Although this is a smaller proportion than for RNA–protein analysis, the number of genes used in the clustering is significantly smaller for RPPA (48 versus 4,291 for RNA). **e**, ESTIMATE tumour purity comparison between mRNA, RNA-seq, and proteome data. ANOVA is used to assess the difference in distribution ($-\log_{10}(P\text{ value})$) of ESTIMATE, stromal, immune, and tumour purity scores across mRNA (microarray), RNA-seq and proteome data. The only significant P value (0.02) is for the cluster 3 stromal score, and higher stromal scores for the proteome drive that difference. **f**, Ischaemia score analysis. Comparison of ischaemia scores of 77 CPTAC tumours, 3 normal samples, and patient-derived xenografts. CPTAC tumours had generally lower ischaemia scores than PDX samples subjected to 30 min of cold ischaemia. Median ischaemia scores are less than 30 min for each subtype and no significant differences were observed across subtypes. Effects due to cold ischaemia therefore appear to be negligible in this CPTAC sample collection.



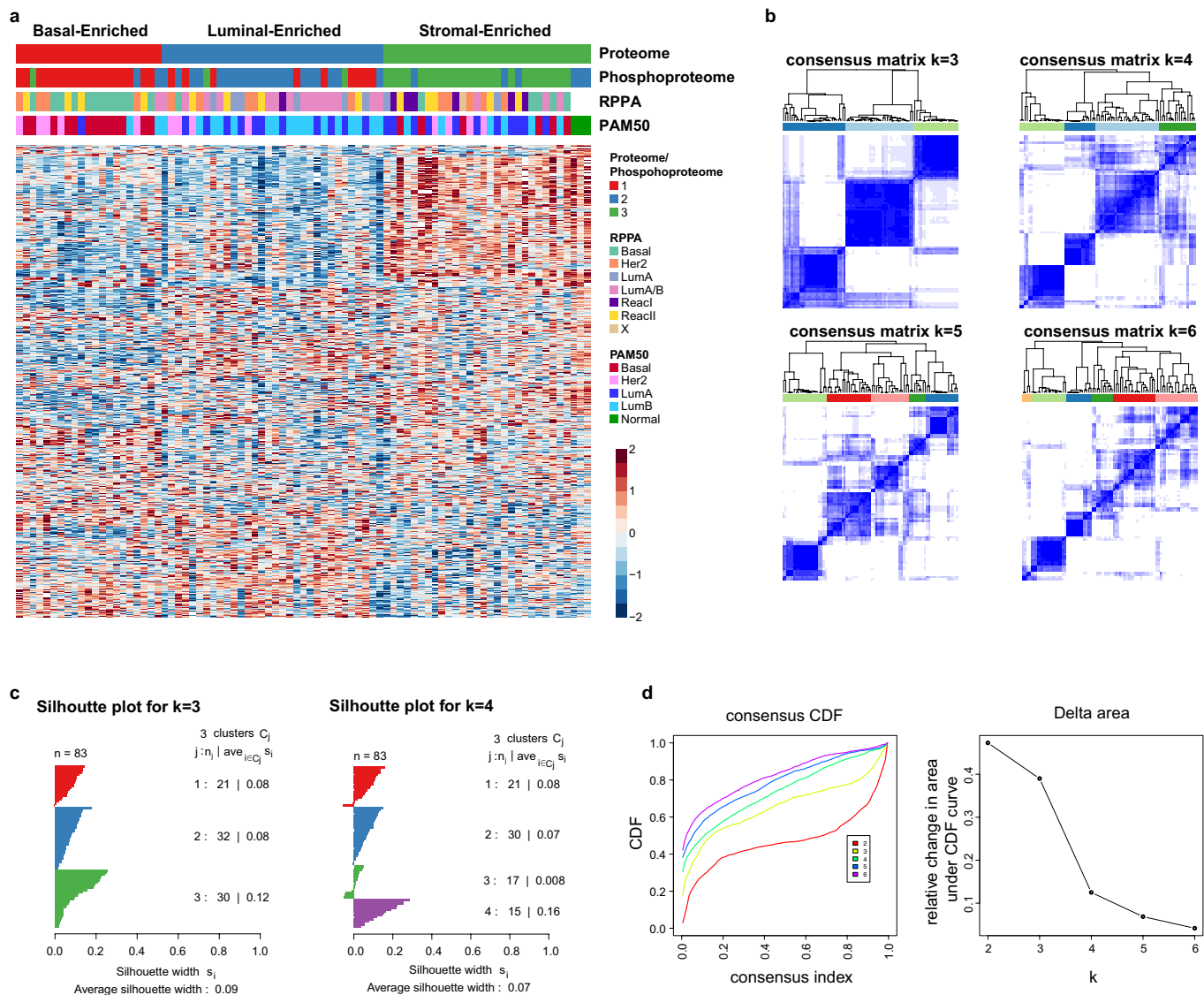
Extended Data Figure 4 | Protein–protein, protein–CNA, and protein–mRNA correlation analyses. **a**, Identification of UBE3A as an E3 ubiquitin ligase that negatively correlates to p53 on the protein level. Pearson correlation and Benjamini–Hochberg-corrected *P* value are shown. **b**, Analysis of counter-regulated genes with negative correlation of

CNA–RNA as well as CNA–protein levels. Negative Pearson correlations are shown with Benjamini–Hochberg-corrected *P* values for CNA–protein correlations. Depicted genes have significant negative correlations at $FDR < 0.05$ in the CNA–RNA and CNA–protein analyses. **c**, Global mRNA–protein correlation and gene set enrichment analysis.



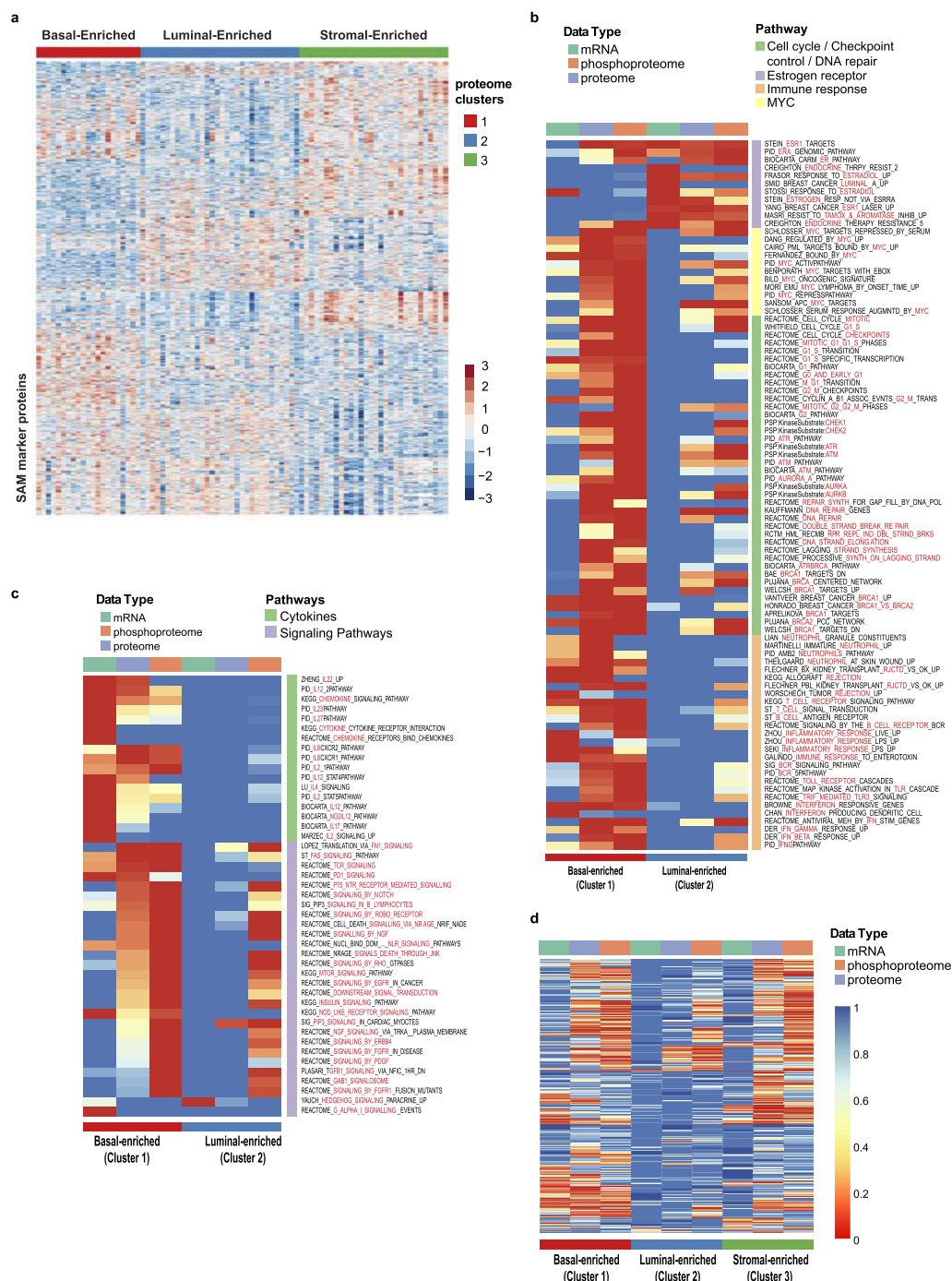
Extended Data Figure 5 | Global CNA effects and comparison of CNA *trans*-effects to knockdown signatures in the LINCIS database. **a**, CNA landscape in the CPTAC tumour collection. The segment-based CNAs of 77 samples were downloaded from TCGA Firehose, including 18 Basal, 12 Her2, 23 Luminal A and 24 Luminal B subtypes. Copy number amplifications were marked in red and deletions in blue. The bottom colour key represents the log₂-transformed copy number value, with CNA = 2 centred at 0. Specific CNA events are seen for chromosome 5q and 10p regions in basal-like tumours. **b**, Correlations of copy number alterations (*x* axis) to phosphoprotein levels (*y* axis) highlight new CNA *cis*- and *trans*-effects. Significant (FDR < 0.05) positive (red) and negative (green) correlations between CNA and phosphoproteins are indicated. Histograms show the fraction (%) of significant CNA *trans*-effects for

each CNA gene. **c**, LINCIS CMap analysis facilitates identification of novel functional candidates for CNA *trans*-effects. Knockdown profiles were compared with CNA–protein *trans*-effects for 502 genes. Genes with a connectivity score > |90| were considered connected and significant *cis*-effects were annotated at an FDR < 0.05. **d**, Basal-like tumour-specific CNAs are candidate regulatory events for EGFR and SRC expression levels. Oncogenic kinases with significant CNA–protein *trans*-effects (left panel), that were regulated in LINCIS short hairpin RNA experiments (right panel; 4 cell lines) and directly measured as LINCIS landmark genes, are shown alongside candidate regulatory genes *CETN3* and *SKP1*. Clinical ER, PR, and HER2 annotation and PAM50 classification are shown in the header rows of each column.



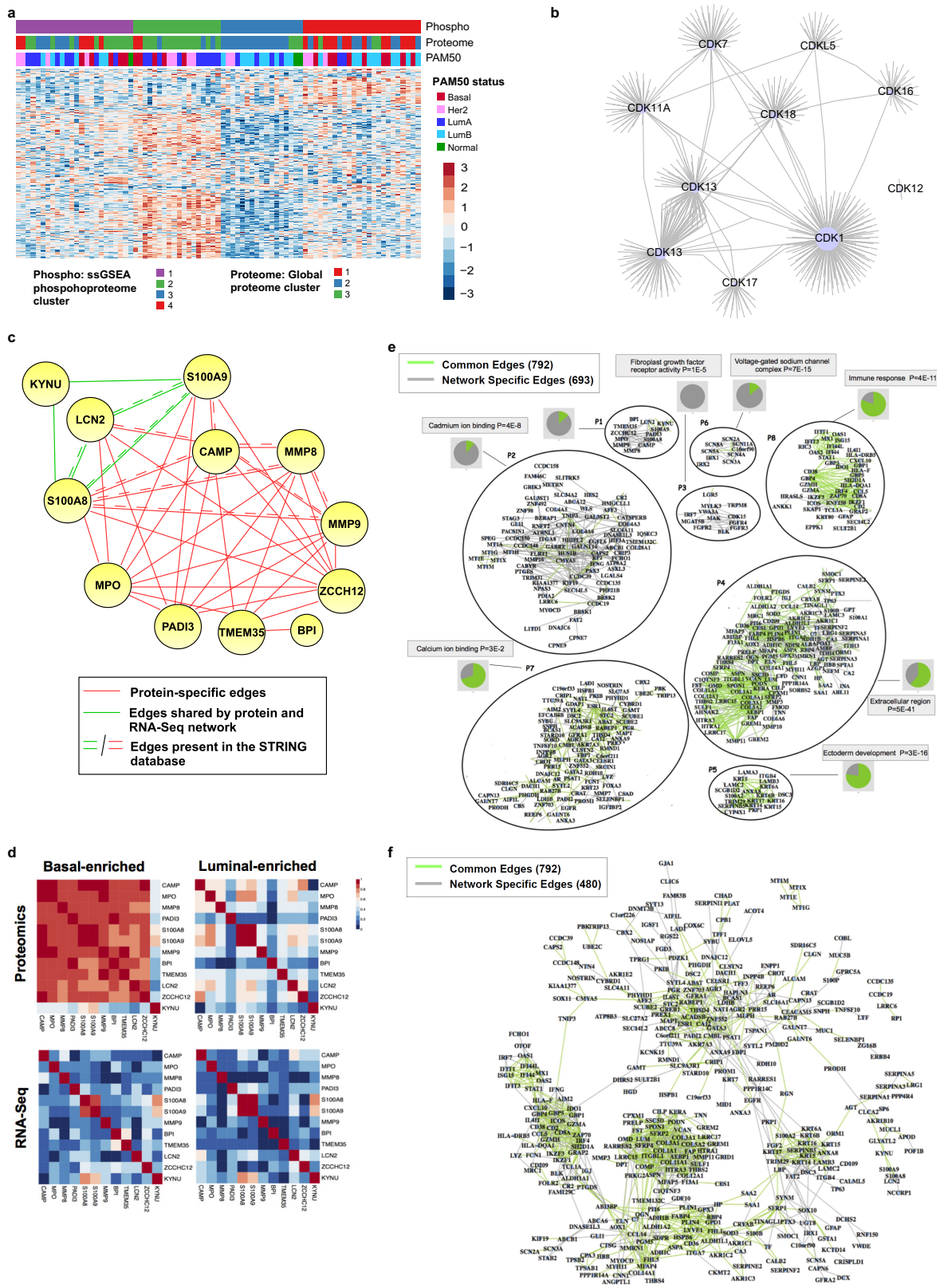
Extended Data Figure 6 | Proteome cluster heat map and stability analysis. **a**, K-means consensus clustering of proteome and phosphoproteome data identifies three subgroups: basal-enriched, luminal-enriched, and stromal-enriched. The heat map represents all 1,521 proteins used for clustering (data set G8). **b**, Identification of optimal proteome clusters for quality-control-passed CPTAC breast cancer tumours. Proteome clusters were derived using consensus clustering based on 1,000 resampled data sets, exploring the range of 2 to 6 K-means clusters. Visualization of consensus matrices from K-means consensus

clustering for $K = 3, 4, 5$ and 6 target clusters. Consensus clustering was performed on 1,521 proteins with no missing values and s.d. > 1.5 . **c**, Silhouette plots were generated to evaluate the coherence of the clustering. Silhouette plots for $K = 3$ and $K = 4$ clusters showing a cleaner separation of clusters for $K = 3$. **d**, On the basis of both visual inspection of the consensus matrix and the delta plot assessing change in consensus cumulative distribution function (CDF) area, three robustly segregated groups were observed. Consensus CDF and delta area (change in CDF area) plots for 2–6 clusters.



Extended Data Figure 7 | Proteome cluster markers and enriched pathways. **a**, Markers (based on SAM analysis; $FDR < 0.01$) discriminate between proteome clusters 1, 2 and 3 (compare to heat map of proteins used to derive clusters depicted in Extended Data Fig. 6a). **b**, Applying a Fisher-exact-test-based enrichment analysis to the proteome, phosphoproteome and mRNA data, gene sets from MSigDB were identified that were unique for each proteome cluster. Heat map showing specific pathways comprising dominant biological themes that are significantly differential by enrichment analysis between basal-enriched and luminal-enriched tumours (Fisher exact test Benjamini–Hochberg-corrected P values are shown; enrichment test performed on marker sets identified using SAM analysis; see Methods; compare to Fig. 3c). **c**, Heat map showing a selection of gene sets significant in basal-enriched

or luminal-enriched tumours exclusively by mRNA, protein or phosphoprotein expression. Cytokine signatures, for example, were strongly captured at the mRNA level, but were seen to only a limited degree at the global protein level, probably because of their typically low protein abundance. By contrast, the vast majority of significant gene sets annotated as ‘signaling’ were enriched only at the phosphoprotein level. **d**, Global heat map representing all gene sets significantly enriched in at least one of the proteomic breast cancer subtypes. The stromal-enriched group was characterized by breast cancer normal-like, adipocyte differentiation, smooth muscle, toll-like receptor signalling and endothelial gene sets, supporting the clustering-based annotation of high stromal and/or adipose content in these tumours (see Supplementary Table 13).

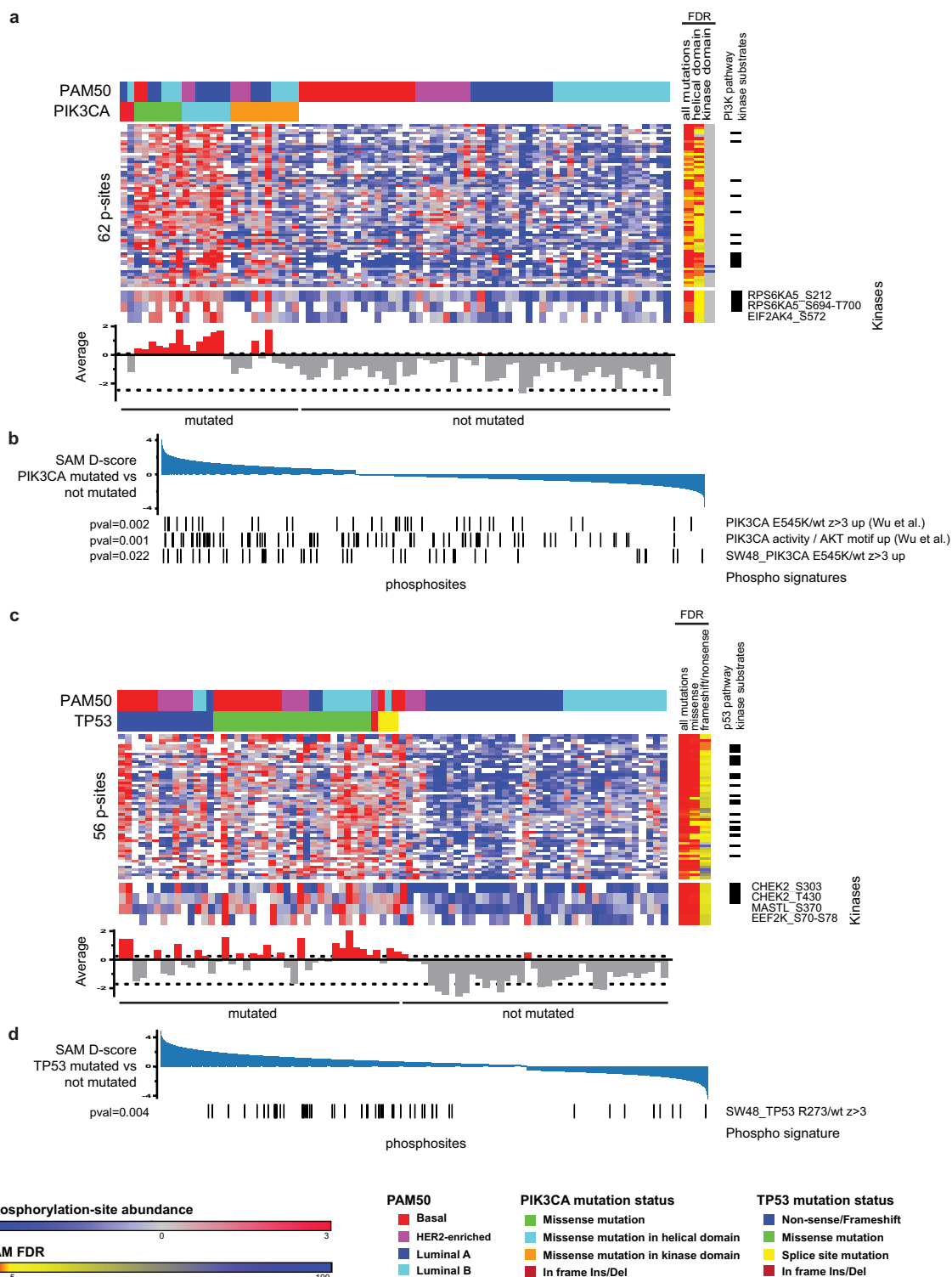


Extended Data Figure 8 | See next page for caption.

Extended Data Figure 8 | Phosphoproteome pathway clustering, kinase-phosphosite multivariate regression, and protein co-expression networks.

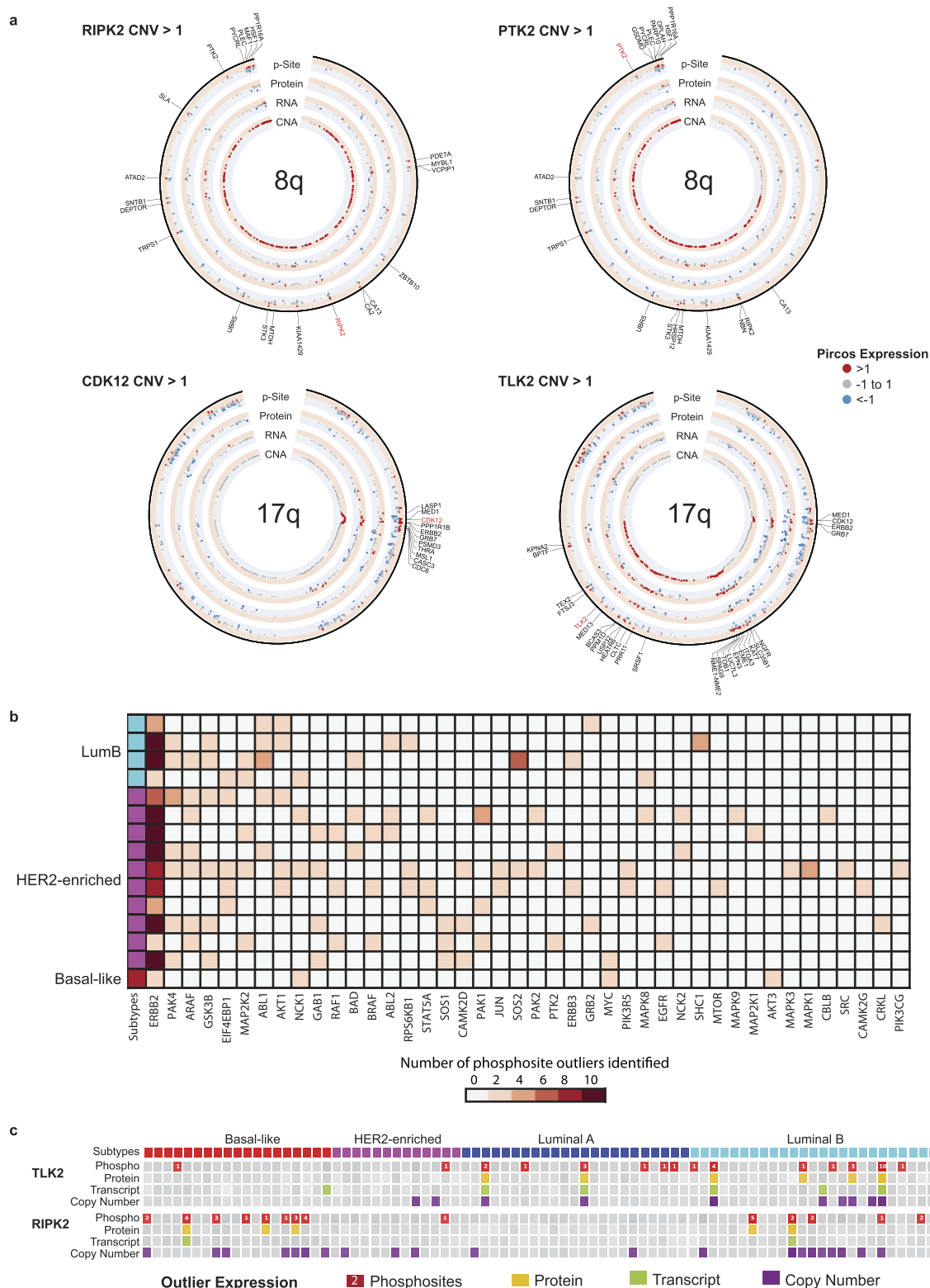
a, Phosphoproteome pathway clustering. Using phosphorylation state as a proxy for activity, deep phosphoproteome profiling allows development of a breast cancer molecular taxonomy on the basis of signalling pathways. *K*-means consensus clustering was performed on pathways derived from single sample GSEA analysis of phosphopeptide data (908 pathways shown). Of four robustly segregated groups, subgroups 2 and 3 substantially recapitulated the stromal- and luminal-enriched proteomic subgroups, respectively. Subgroup 4 included a significant majority of tumours from the basal-enriched proteomic subgroup, but was admixed particularly with luminal-enriched samples. This subgroup was defined by high levels of cell cycle and checkpoint activity. All basal and a majority of non-basal samples in this subgroup had *TP53* mutations. Subgroup 1 was a novel subgroup defined exclusively in the phosphoproteome pathway activity domain, with no enrichment for either proteomic or PAM50 subtypes. It was defined by G protein, G-protein-coupled receptor, and inositol phosphate metabolism signatures, as well as ionotropic glutamate signalling. **b**, Analysis of the regulatory relationship between outlier kinases (see Supplementary Table 19) and phosphopeptides by regulatory multivariate regression analysis (see Methods) identified CDK1 as the most highly connected of the outlier cyclin-dependent kinases, with highest centrality (based on node-degree; see Methods) among the outlier CDKs and seventh highest centrality among all the outlier kinases considered in the remMap analysis. Each line represents a phosphosite-kinase relationship. **c–f**, Analysis of differences in the co-expression patterns among genes/proteins across different subgroups. A Joint Random Forest method was applied to simultaneously build gene co-expression and protein

co-expression networks (Supplementary Table 17, and Methods). Modules in these networks revealed different interaction patterns between basal-enriched and luminal-enriched subgroups. **c**, Network module P1 of the protein co-expression network, defined chiefly in the proteome space. This module contained 12 genes connected by 39 edges, among which 34 were protein-specific and 5 were shared by both the protein and mRNA co-expression networks. Many edges were supported by published information and were contained in the STRING database. Edges in red are specific to the protein co-expression network; edges in green are shared by both protein and gene co-expression networks; edges indicated by double lines are contained in the STRING database with confidence score greater than 0.15. MMP9, one of the central proteins in this module, contributes to metastatic progression and is a potential target for anti-metastatic therapies for basal-like/triple-negative breast cancer. **d**, Heat maps of the absolute correlation across each pair of genes in module P1 (shown in **c**), based on either protein or gene expression data for samples in the basal-enriched and luminal-enriched subgroups, respectively. The MMP9 protein was strongly co-expressed with the other members of the module only in the basal-enriched subgroup. Notably, this observation is dependent on protein data; the correlation at the mRNA level for this module was consistently low in both the basal-enriched and luminal-enriched subgroups indicating that these events coherently occur at the proteomic level. **e**, Co-expression network based on proteomics data. The network contains 693 proteomic network-specific edges (grey) and 792 edges shared with the RNA-seq network (green). For each module, the most enriched category and corresponding Benjamini–Hochberg-adjusted *P* value is reported. Pie charts adjacent to each module show the proportion of proteomics-specific edges (grey area) and edges shared between proteomics and RNA-seq data (green area). **f**, RNA-seq network.



Extended Data Figure 9 | Phosphoproteome signatures of *PIK3CA*- and *TP53*-mutated tumours highlight activated key regulators and indicate frequency of activation. **a, c,** Phosphosites upregulated in mutated tumours (SAM FDR < 0.05 across all tumours and independently also across luminal tumours; average phosphosite signal for all markers shown as bar graph). To avoid confounding by intrinsic subtype-specific distinctions, only markers that were significantly identified both in analyses covering all tumours and analyses restricted to luminal tumours were selected (FDR < 0.05). Colour bars in the margins indicate FDRs for grouped analysis of different mutation classes and indicate kinase substrates of known kinases in the respective pathways. Significantly regulated kinase phosphosites are annotated. The average phosphorylation signal of the marker phosphosites provides a read-out for PI3K and TP53 pathway activity in mutated tumours (histogram below heat map). A 95% prediction confidence interval (indicated by dashed lines) across the average signal in non-mutated tumours was chosen in order

to discriminate active from non-active tumours. The most strongly activated *PIK3CA* kinase domain mutant tumour differed from the other nine kinase domain mutant tumours, as it contained an amino acid side chain charge neutral H1047L instead of the more common positively charged H1047R mutation. Among the 62 phosphosites identified that were significantly upregulated in *PIK3CA*-mutated tumours, 13 phosphosites were found on phosphoproteins that are known substrates of well-annotated kinases in the PIK3CA pathway (**a**, right column). In the mutant *TP53* analysis, a total 20 phosphosites were found on phosphoproteins that are known substrates of well annotated kinases in the p53 pathway (**c**, right column). **b, d,** Upregulated phosphosite sets were derived from isogenic *PIK3CA* and *TP53* mutant versus wild-type cell-line pairs and tested for enrichment within mutant versus wild-type CPTAC tumours using single sample GSEA. Significantly enriched phosphosite sets are shown ($P < 0.05$).



Extended Data Figure 10 | Pircos plots, kinase outliers and outliers in the ERBB2 pathway. **a**, Pircos (proteogenomics circo) plots for 8q and 17q showing median CNA, RNA, protein, and phosphosite expression for 20 tumours with amplification in 8q based on *RIPK2* CNA > 1; 23 tumours with amplification in 8q based on *PTK2* CNA > 1; 15 tumours with amplification in 17q based on *CDK12* CNA > 1; and 10 tumours with amplification in 17q based on *TLK2* CNA > 1. Red indicates expression > 1, blue < -1, and grey between -1 and 1. Genes with both copy number amplification (CNA > 1) and increased phosphosite expression (p-site > 1) are labelled. **b**, Phosphosite outliers in known *ERBB2* signalling genes. To better understand the downstream effects of *ERBB2* amplification,

phosphosite outliers in known *ERBB2* signalling genes (MSigDB' pathway set, 'KEGG_ERBB_SIGNALING_PATHWAY') were identified for the 15 samples that had *ERBB2* phosphosite outlier status. Forty-one genes were identified as having a phosphosite outlier in at least one of the *ERBB2*-amplified samples. *PAK4* and *ARAF* phosphosite outlier status were found in seven of the 15 *ERBB2* kinase outlier samples; *GSK3B* outliers were found in 6 samples; and *EIF4EBP1*, *MAP2K2*, *ABL1* and *AKT1* outlier status was found in 5 of the 15 samples. **c**, Proteogenomic outlier expression analysis for *TLK2* and *RIPK2*. Samples with outlier phosphosite (red), protein (yellow), RNA (green) and copy number (purple) expression are shown. Phosphosite squares indicate per-sample outlier phosphosites.

Activation of NMDA receptors and the mechanism of inhibition by ifenprodil

Nami Tajima¹, Erkan Karakas¹, Timothy Grant², Noriko Simorowski¹, Ruben Diaz-Avalos², Nikolaus Grigorieff² & Hiro Furukawa¹

The physiology of *N*-methyl-D-aspartate (NMDA) receptors is fundamental to brain development and function. NMDA receptors are ionotropic glutamate receptors that function as heterotetramers composed mainly of GluN1 and GluN2 subunits. Activation of NMDA receptors requires binding of neurotransmitter agonists to a ligand-binding domain (LBD) and structural rearrangement of an amino-terminal domain (ATD). Recent crystal structures of GluN1–GluN2B NMDA receptors bound to agonists and an allosteric inhibitor, ifenprodil, represent the allosterically inhibited state. However, how the ATD and LBD move to activate the NMDA receptor ion channel remains unclear. Here we applied X-ray crystallography, single-particle electron cryomicroscopy and electrophysiology to rat NMDA receptors to show that, in the absence of ifenprodil, the bi-lobed structure of GluN2 ATD adopts an open conformation accompanied by rearrangement of the GluN1–GluN2 ATD heterodimeric interface, altering subunit orientation in the ATD and LBD and forming an active receptor conformation that gates the ion channel.

NMDA receptors are critically involved in brain development and function, including learning and memory formation. NMDA receptors belong to the family of ionotropic glutamate receptors, which are glutamate-gated ion channels comprised of three major families, α -amino-3-hydroxy-5-methyl-4-isoxazole propionic acid (AMPA) (GluA1–4), kainate (GluK1–5), and NMDA receptors (GluN1, GluN2A–D, and GluN3A, B)¹. NMDA receptors are obligatory heterotetramers mainly composed of two copies each of the GluN1 and GluN2 subunits, which bind glycine and L-glutamate, respectively. Under physiological conditions, the opening of the NMDA receptor ion channel requires concurrent binding of glycine and L-glutamate^{2–4}, and relief of magnesium block at the ion channel pore by membrane depolarization^{5,6}. The resulting calcium flux⁷ triggers a cascade of signal transduction necessary for synaptic plasticity⁸. Dysfunctional NMDA receptors are implicated in various neurological diseases and disorders such as Alzheimer's disease, depression, stroke, epilepsy and schizophrenia^{1,9}.

NMDA receptor subunits, like those of other ionotropic glutamate receptor family members, are composed of multiple domains including an ATD, LBD, transmembrane domain (TMD) and carboxy-terminal domain (CTD) (Extended Data Fig. 1). Binding of neurotransmitter agonists to the LBD produces a large conformational change involving closure of the bi-lobed structure that is required for ion channel gating in all ionotropic glutamate receptors^{10–12}, but a distinctive feature of NMDA receptors is that activity is also robustly regulated by the ATD¹³. For example, the ATD controls the open probability and speed of deactivation^{14,15}, and binds allosteric modulator compounds to regulate ion channel activity¹⁶. In contrast to NMDA receptors, there is no apparent role for the ATDs of AMPA and kainate receptors^{17–20} in regulating the ion channel activities, even though they are essential for subunit assembly²¹. The recent crystal structures of intact heterotetrameric GluN1–GluN2B NMDA receptors complexed with agonists and allosteric inhibitors, ifenprodil or Ro 25-6981, revealed that the ATD and LBD interact tightly via a large interface area, unlike GluA2 AMPA receptor and GluK2 kainate receptor whose ATDs and LBDs interact minimally^{17,20,22,23}, implying that activation of NMDA receptors requires concerted conformational alterations in the ATD

and LBD^{19,22,23}. The structures of the intact GluN1–GluN2B NMDA receptors^{22,23} and of the isolated ATDs complexed to ifenprodil²⁴ or zinc²⁵ showed a closed conformation of the bi-lobed GluN2B ATD architecture^{22–25}, probably representing the 'allosterically inhibited' functional state. In the presence of agonists, NMDA receptors are known to reside in active states that can trigger ion channel opening, as well as desensitized states with a channel that is closed even in the presence of bound agonists²⁶. Despite accumulating structural information on intact NMDA receptors^{22,23}, as well as isolated ATDs^{24,25,27} and LBDs^{11,28,29}, there is a lack of structures representing the active state and the mechanism of activation has remained unclear. In this study, we present structures of the isolated ATD in the apo-state and of the intact receptor in the activated conformation, providing a detailed mechanistic picture of receptor activation.

Opening of the GluN2B ATD and subunit rearrangement

The only available structures for the heterodimeric NMDA receptor ATDs to date are those bound to allosteric inhibitors ifenprodil and Ro 25-6981, representing the allosterically inhibited state^{22–25}. We reasoned that by conducting structural studies without allosteric inhibitors, we could capture the ATD conformation that can activate the NMDA receptor ion channel. Thus, we determined the crystal structure of GluN1–GluN2B ATDs in the absence of an allosteric inhibitor (apo-GluN1b–GluN2B ATD) at 2.9 Å resolution (Extended Data Table 1). We crystallized the purified GluN1b–GluN2B ATD proteins complexed to a Fab fragment derived from mouse monoclonal IgG to improve the quality of the crystals (Extended Data Fig. 2). The crystallographic analysis shows heterodimeric GluN1–GluN2B ATDs that have a bi-lobed architecture composed of the regions previously called R1 and R2 in the structure of GluN1b–GluN2B ATD bound to the allosteric inhibitor ifenprodil²⁴ (Fig. 1). There are a number of differences between the structures of the apo-GluN1b–GluN2B ATD and the ifenprodil-bound GluN1b–GluN2B ATD²⁴. The most apparent difference is the separation of GluN1b R1 and GluN2B R2 in the apo-GluN1b–GluN2B ATD, owing to the $\sim 20^\circ$ rigid-body opening of the GluN2B ATD bi-lobed structure in the apo-GluN1b–GluN2B ATD compared to that in the ifenprodil–GluN1b–GluN2B ATD (Fig. 1d). This observation is

¹Cold Spring Harbor Laboratory, W. M. Keck Structural Biology Laboratory, Cold Spring Harbor, New York 11724, USA. ²Janelia Research Campus, Howard Hughes Medical Institute, Ashburn, Virginia 20147, USA.

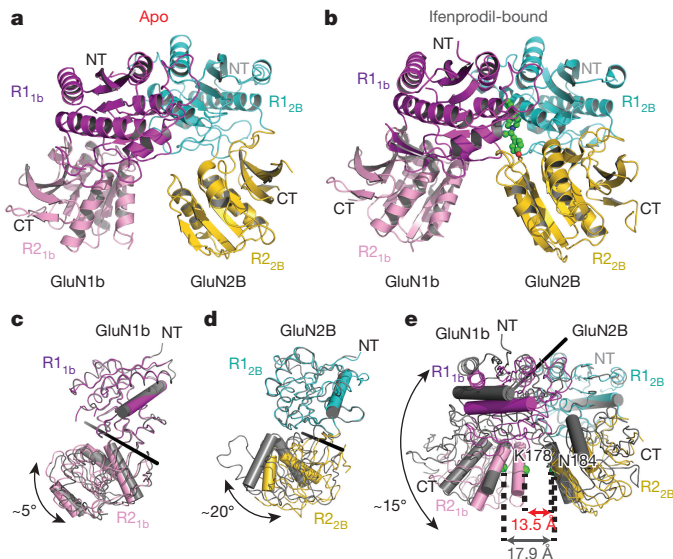


Figure 1 | Structures of GluN1b–GluN2B ATD heterodimers.

a, b, Crystal structure of the GluN1b–GluN2B ATD heterodimer in the apo state solved at 2.9 Å (**a**) in comparison with the ifenprodil-bound structure (PDB ID, 3QEL) (**b**). The R1 and R2 lobes are coloured magenta and light pink for GluN1b ATD, and cyan and yellow for GluN2B ATD. Ifenprodil is represented by green spheres. **c, d,** Superimposition of the R1 lobes of GluN1b (**c**) and GluN2B (**d**) in the apo and ifenprodil-bound (grey) forms illustrates the relative 'opening' between R1 and R2 lobes. **e,** Superimposing the GluN2B R1 lobes of apo and ifenprodil-bound forms reveals an ~15° rotation of GluN1b ATD relative to GluN2B ATD along the axis of rotation (black rod). The distance of the R2 lobes in the GluN1b–GluN2B heterodimers is measured between GluN1b Lys178 and GluN2B Asn184 (green spheres). CT, C terminal; NT, N terminal.

consistent with previous work suggesting that GluN2B ATD has open-cleft and closed-cleft conformations in the absence and presence of ifenprodil, respectively, on the basis of luminescence resonance energy transfer studies³⁰. Another major difference is the rearrangement of the GluN1b and GluN2B subunits involving an ~15° rotation relative to one another (Fig. 1e). This rearrangement brings the lower lobes (R2) of GluN1–GluN2B considerably closer together in the apo-GluN1b–GluN2B ATD compared with the ifenprodil–GluN1b–GluN2B ATD (Fig. 1e, Extended Data Fig. 2c). For example, the distance between the Cα atoms of GluN1b Lys178 and GluN2B Asn184 in apo-GluN1b–GluN2B ATD is 4.4 Å closer than in the ifenprodil–GluN1b–GluN2B ATD (Fig. 1e).

As the subunit arrangement in the apo-GluN1b–GluN2B ATD in our crystal structure is different from that previously observed in the ifenprodil–GluN1b–GluN2B ATD²², we sought to validate its physiological relevance. Towards this end, we tested whether an inter-subunit disulfide bond can form at the subunit interface observed in the apo-GluN1b–GluN2B ATD, but not in the ifenprodil–GluN1b–GluN2B ATD in the context of the intact GluN1–GluN2B NMDA receptor by mutating GluN1 and GluN2B residues that are proximal to each other. We expected a spontaneous disulfide bond to form between the mutated cysteines in the intact GluN1–GluN2B NMDA receptor if the subunit interface observed in the crystal structure is physiological. We engineered cysteine residues at GluN1b Phe113 and GluN2B Ala107, and at GluN1b Gly331 and GluN2B Glu75, expressed and purified the mutant GluN1b–GluN2B NMDA receptor in the context of the intact ion channel, and conducted western blot analysis under non-reducing conditions to detect band shifts (Extended Data Fig. 3a). In the two selected positions, the disulfide bonds are formed only when the cysteine mutant (Extended Data Fig. 3) of GluN1 and that of GluN2B are co-expressed and detected by an anti-GluN1 and anti-GluN2B western blot in the absence of

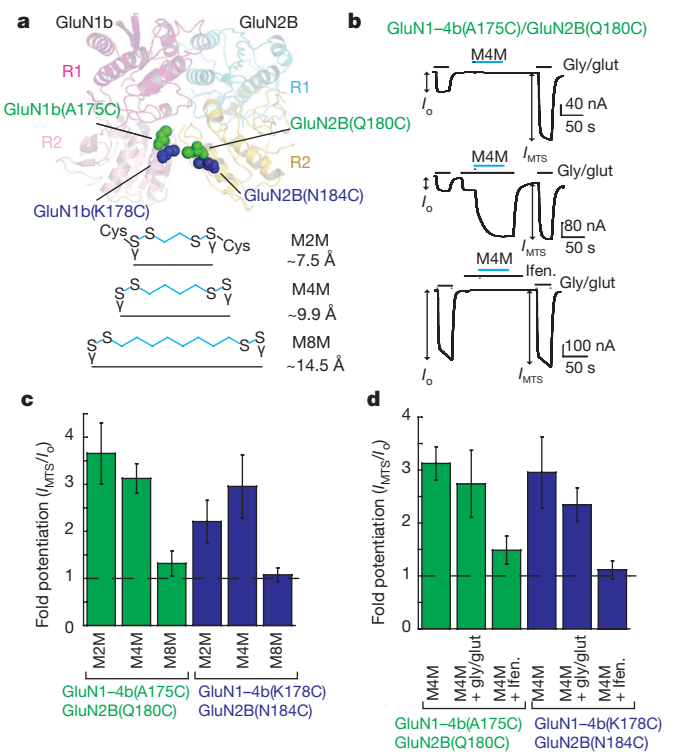


Figure 2 | Conformational trap identifies the apo-GluN1b–GluN2B ATD structure as the 'active' form.

a, Location of engineered cysteines in the crystal structure of the apo-GluN1b–GluN2B ATD (GluN1–4b(Ala175Cys)/GluN2B(Gln180Cys) in green spheres and GluN1–4b(Lys178Cys)/GluN2B(Asn184Cys) in blue spheres). **b,** Application of 200 μM M4M in the presence or absence of 100 μM agonists (glycine and glutamate (gly/glut)) potentiates the macroscopic current measured at the holding potential of –60 mV by two-electrode voltage clamp. No potentiation was observed when M4M was applied in the presence of ifenprodil (ifen.). Shown here are the representative recording profiles for the GluN1–4b(Ala175Cys)/GluN2B(Gln180Cys) pair. **c, d,** Fold of potentiation is presented as I_{MTS}/I_0 (I , current amplitude) as measured in **b** for bifunctional MTS with different linker lengths (**c**) and M4M applied in different functional states (**d**). Error bars represent ±s.d. for data obtained from at least five different oocytes per experiment.

β-mercaptoethanol. When the cysteine mutants of one subunit is co-expressed with the wild type of the other subunit, no disulfide bonds are formed, indicating that they are specifically formed by the engineered cysteines. Taken together, the above experiments show that the GluN1–GluN2B subunit arrangement observed in the apo-GluN1b–GluN2B ATD crystal structure exists in the context of the intact GluN1b–GluN2B NMDA receptor.

Active conformation of the ATD

To understand the functional state that the crystal structure of the apo-GluN1b–GluN2B ATD may represent, we next attempted to stabilize the conformation observed in the crystal structure and assessed the ion channel activity. We engineered cysteines at the positions in the lower lobes (R2) of the GluN1b and GluN2B ATDs (GluN1b(Ala175Cys)/GluN2B(Gln180Cys) and GluN1b(Lys178Cys)/GluN2B(Asn184Cys)), which face each other and should 'trap' the conformation observed in the crystal structure by tethering the engineered cysteines with bifunctional methanthiosulfonate (bi-MTS) reagents (Fig. 2). The distances between the mutated residues are closer in apo-GluN1b–GluN2B ATD than in ifenprodil–GluN1b–GluN2B ATD as mentioned above (Fig. 1e). When bi-MTS, equal or shorter in length than M4M, binds to the lower lobes of the GluN1b–GluN2B heterodimers, we reasoned that the conformation observed in the apo-GluN1b–GluN2B ATD with the open GluN2B bi-lobed architecture and the rearranged GluN1–GluN2B

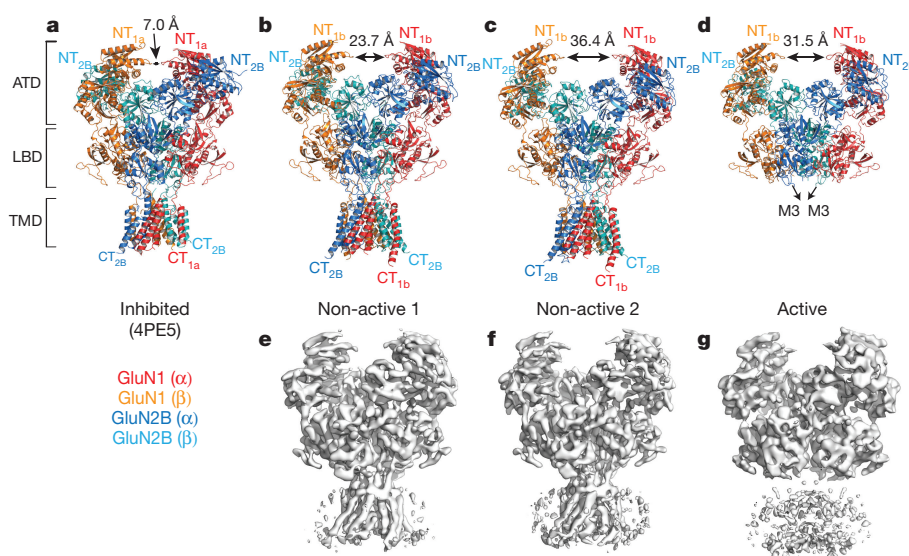


Figure 3 | Overall structures of the intact GluN1–GluN2B NMDA receptors at different conformational states. **a**, The crystal structure of GluN1a–GluN2B NMDA receptor in complex with glycine, L-glutamate and ifenprodil (PDB ID, 4PE5). **b–d**, Cryo-EM structures of glycine- and L-glutamate-bound GluN1b–GluN2B NMDA receptors classified to reveal different conformations representing the ‘non-active’ (**b**, **c**) and ‘active’ (**d**) states. Subunits: red, GluN1 (α); orange, GluN1 (β); blue, GluN2B (α); cyan, GluN2B (β).

cyan, GluN2B (β). The two ‘non-active’ 3D classes (non-active 1 and 2) have different distances between the two GluN1–GluN2B ATD heterodimers represented as the distance between the C α atoms of Glu320 (299 in GluN1a) of GluN1b (α) and GluN1b (β) (double-sided arrows). The amino and carboxy termini and approximate domain boundaries are indicated. **e–g**, The cryo-EM maps for the ‘non-active 1’ (**e**), ‘non-active 2’ (**f**) and ‘active’ (**g**) states.

subunit orientation should be trapped. To test this, we co-expressed the cysteine mutants of GluN1b and GluN2B in *Xenopus* oocytes and probed the effect of the bi-MTS reagents on the macroscopic current of NMDA receptor by two-electrode voltage clamp. We initialized this experiment by testing bi-MTS with the four-carbon linker (M4M in Fig. 2a) as the estimated distances between the γ -sulfur atom of the mutated cysteines in the GluN1b(Ala175Cys)/GluN2B(Gln180Cys) and GluN1b(Lys178Cys)/GluN2B(Asn184Cys) mutants of apo-GluN1b–GluN2B ATD are ~ 10 Å and ~ 9 Å, respectively, roughly matching the length of M4M. The application of M4M to the GluN1b(Ala175Cys)/GluN2B(Gln180Cys) and GluN1b(Lys178Cys)/GluN2B(Asn184Cys) mutants potentiates the NMDA receptor currents by ~ 3 – 4 -fold (Fig. 2b, c, Extended Data Fig. 4). No such effect is observed when the cysteine mutants of one subunit are co-expressed with the wild type of the other subunit, indicating that the observed functional effect is specific to the engineered cysteines (Fig. 2a and Extended Data Fig. 5a, b). We suggest that this potentiating effect by the bi-MTS conformational trap favoured the NMDA receptor ion channel to reside in the ‘active’ form. The effect of M4M is observed both in the presence and absence of glycine and glutamate, indicating that conformational alteration in the ATD is independent of agonist binding in the LBD. Furthermore, the potentiation effect was also observed when M2M was applied to both of the above mutant pairs, indicating that the GluN1b–GluN2B distance in R2 may move even closer than observed in the crystal structure, consistent with the single-particle electron cryomicroscopy (cryo-EM) structures shown later. By contrast, when adding M8M, a bi-MTS agent that is 4–5 Å longer than the inter-cysteine distances observed in the apo-GluN1b–GluN2B ATD, no potentiating effect was observed, supporting the view that the distance between the R2 lobes of GluN1b–GluN2B must be reduced during activation (Fig. 2c, Extended Data Fig. 5). Finally, when M4M was applied in the presence of ifenprodil, we observe little or no potentiating effect indicating that it traps the active conformation of GluN1b–GluN2B ATDs but not the inhibited conformation as represented by the crystal structure of the ifenprodil–GluN1b–GluN2B ATD (Fig. 2b, d). Taken together, these experiments indicate that the protein conformation observed in the crystal structure of the apo-GluN1b–GluN2B ATD probably represents the active conformation that facilitates ion channel opening.

Structures of intact GluN1b–GluN2B NMDA receptors

We next investigated how the changes in the GluN1–GluN2B ATD conformation alter subunit arrangement and inter-ATD–LBD interactions to ultimately mediate gating of the ion channel. To answer this, we determined cryo-EM structures of the intact heterotetrameric rat GluN1b–GluN2B NMDA receptor ion channel in the presence of glycine and L-glutamate and in the absence of ifenprodil. The cryo-EM structures were reconstructed at resolutions better than 7 Å and revealed clear secondary structure elements (Fig. 3, Extended Data Figs 6, 7 and Extended Data Table 2). The cryo-EM structures show conservation of general features observed in the recent full-length NMDA receptor crystal structures, including a dimer of GluN1–GluN2B heterodimers arrangement at the ATD and LBD layers, the domain swap between the ATD and LBD, and pseudo-four-fold symmetrical subunit arrangement at the TMD^{22,23}. Importantly, three-dimensional (3D) classification of the cryo-EM data revealed different conformational states present in the data set (Fig. 3). Overall, there are roughly three distinct conformations, which we define as ‘non-active 1’, ‘non-active 2’, and ‘active’ (Fig. 3). When compared to the crystal structure of the intact NMDA receptors bound to ifenprodil, glycine and L-glutamate^{22,23}, which represent the allosterically inhibited functional state, all of the 3D classes contain a GluN2B ATD open bi-lobed architecture, with an $\sim 14^\circ$ – 21° opening similar to the crystal structure of the apo-GluN1b–GluN2B ATD. This opening of the GluN2B ATD increases the distance between the two GluN1 ATDs by as much as ~ 29 Å in the intact NMDA receptor compared to the ifenprodil-bound form (Fig. 3). The comparison shows that, upon ifenprodil binding, the R1 lobe moves relative to the LBD and TMD to close the bi-lobed architecture of the GluN2B ATD, as well as the gap between the two GluN1 ATDs to inhibit receptor activity.

The two 3D classes, non-active 1 and non-active 2, are both in the state where agonists are bound to the LBD but the ion channel is closed. When focusing on the ATD, both non-active 1 and non-active 2 do not display the $\sim 15^\circ$ rotation of the GluN1b and GluN2B subunits relative to one another as observed in the crystal structure of the apo-GluN1b–GluN2B ATD, which represents a conformation that can activate the receptor. The arrangements of the dimer of the GluN1b–GluN2B ATD

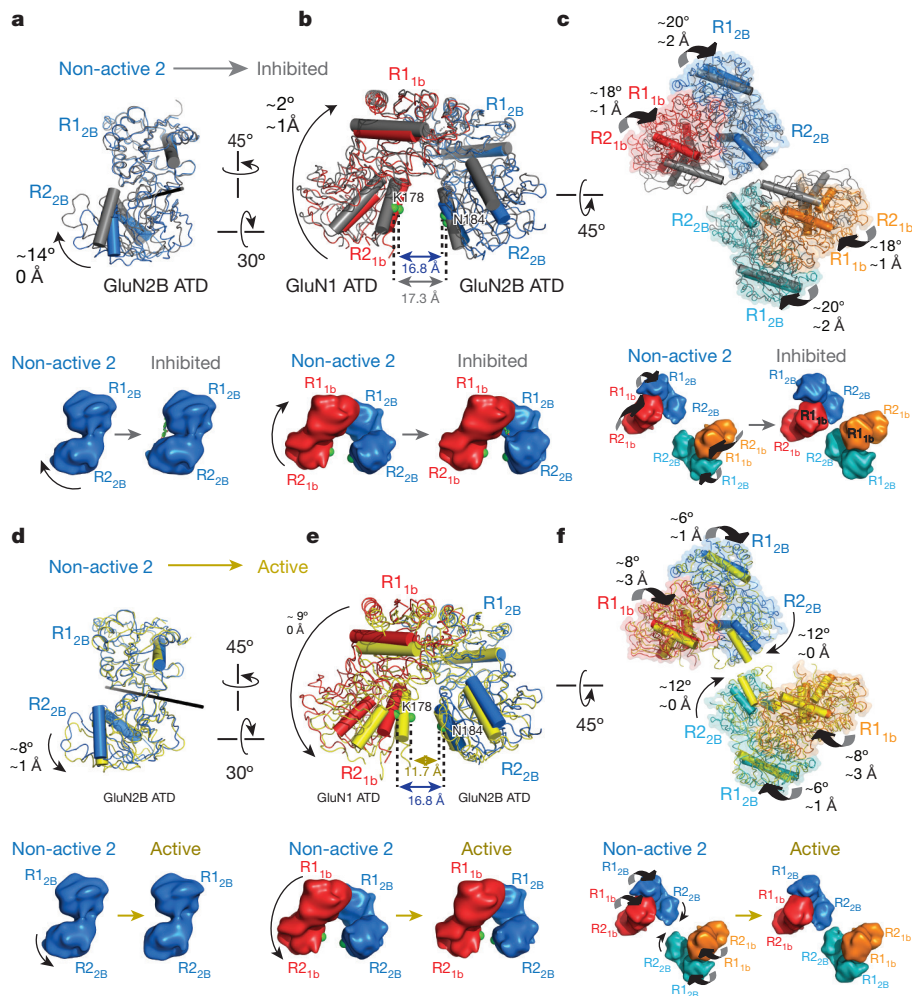


Figure 4 | Changes in conformation and heterotetrameric subunit arrangement in the ATD during ifenprodil inhibition and receptor activation. **a–f**, Comparison of 3D classes show movements (arrowed arcs) of the bi-lobed architecture in the GluN2B ATD (**a**, **d**), and rearrangement of the GluN1b–GluN2B ATD heterodimer (**b**, **e**) and heterotetramer (**c**, **f**) during transition from the agonist-bound ‘non-active 2’ conformation (same colour code as in Fig. 3) to the ifenprodil-bound ‘inhibited’ conformation (PDB ID, 4PE5) (grey) (**a–c**) or to the ‘active’ conformation

(yellow) (**d–f**). Schematic diagrams are shown below each panel. **a**, **d**, Superimposition of GluN2B ATD R1 lobes show relative movement of R2 (around black rods) (**a**, **d**) and rearrangement in the pattern of subunit arrangement (**b**, **e**) in different functional states. **c**, **f**, GluN1b–GluN2B ATD heterotetramer from different 3D classes are compared by aligning the centres of masses (COMs) of the ATD heterotetramer, LBD heterotetramer and individual LBDs. Ifenprodil shown as green sticks.

dimers (Fig. 4c), as well as the dimer of GluN1b–GluN2B LBD dimers (Fig. 5a, c, e) are similar to those observed in the crystal structure of the intact NMDA receptor (‘inhibited’ conformation) (Figs 4 and 5). Consequently, the ion channel pores at the TMD remain closed, confirming that both cryo-EM classes probably represent non-gating or ‘non-active’ conformations. The difference in non-active 1 and non-active 2 is the extent of bi-lobe opening in the GluN2B ATD, where non-active 2 has an $\sim 7^\circ$ more open conformation resulting in ~ 13 Å larger separation between the GluN1 ATDs (Fig. 3). Even though we tentatively call these two conformations ‘non-active’, it remains uncertain whether they represent functional states equivalent to the ‘pre-open’ state observed in non-NMDA receptors^{17,31} or a ‘desensitized’ state.

Active conformation

One of the cryo-EM classes, ‘active’ (Fig. 3), shows the cleft of the bi-lobed GluN2B ATD architecture opened by $\sim 22^\circ$ and a GluN1b–GluN2B heterodimeric subunit rotated by $\sim 12^\circ$, compared to the ifenprodil-bound intact NMDA receptors, which is notably similar to the apo-GluN1b–GluN2B ATD crystal structure representing the active ATD conformation (Figs 1 and 4e and Extended Data Fig. 7). In the heterotetrameric NMDA receptor, the GluN1b–GluN2B

heterodimer pairs rotate by $\sim 12^\circ$ in opposite directions (Fig. 4f). Importantly, this 3D structure of the active conformation of the ATD also shows large differences in the subunit arrangement of LBDs compared to the other 3D classes representing ‘non-active’ ATDs, and is also different from the recent crystal structures of the glycine, L-glutamate and ifenprodil complexes^{22,23}. Specifically, when transitioning from the non-active 2 to active conformation, the two pairs of GluN1b–GluN2B LBD heterodimers rotate by $\sim 13.5^\circ$ (Fig. 5b, d, f). These subunit movements in the LBD cause movement of the residues at the LBD–TMD linkers (Fig. 5). For example, when focusing on the residues located right above the pore formed by the M3 TMD helices (Fig. 6a), the concerted movement between the ATD and LBD going from non-active 2 to active described above causes a vertical movement of GluN1b Arg684 and the lateral separation of GluN2B Glu658 by 7 Å and 11 Å, respectively, to dilate the gating ring, a movement that is likely to lead to ion channel gating³² (Figs 5d, f and 6, Supplementary Videos 1 and 2). Thus, this cryo-EM class is structurally and functionally consistent with an ‘active’ conformation for GluN1–GluN2B NMDA receptors. Although there is clear density for most of the domains in the active conformation of the receptor, the density for the TMD is not resolved in sufficient detail to directly observe opening of the ion channel, as is the case

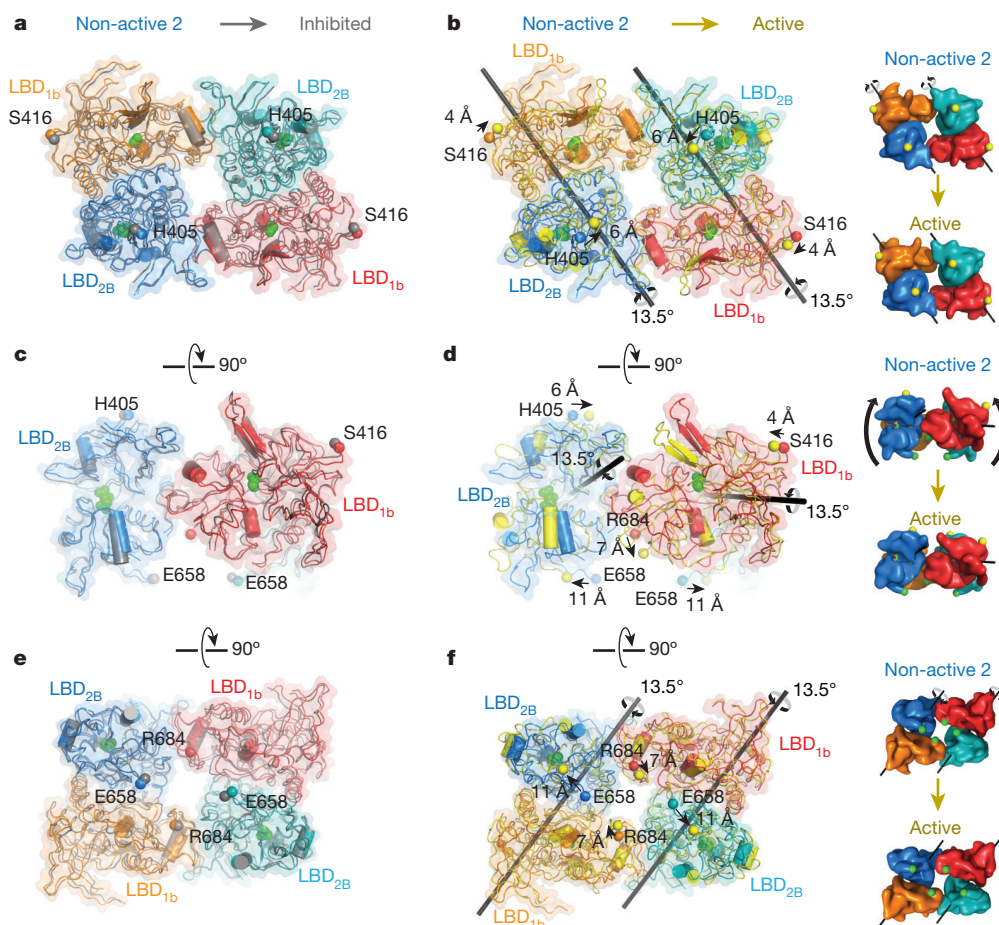


Figure 5 | Conformational changes at the LBD during ifenprodil inhibition and receptor activation. a–f, The same superimposition as in Fig. 4c, f showing the LBD tetramers in the ‘non-active 2’ (same colour code as in Fig. 3), the ‘inhibited’ (grey) (a, c, e) and the ‘active’ (yellow) states (b, d, f) viewed from the ATD (a, b), side (c, d) and TMD (e, f). The LBD heterodimers rotates (around black rods) during transition from

non-active 2 to active, whereas little or no change occurs between non-active 2 and inhibited. Co atoms of the residues at the ATD–LBD linker (GluN1b His405 and GluN2B Ser416) and the LBD–TMD linker (GluN1b Arg 684 and GluN2B Glu 658) are shown as spheres. Glycine and L-glutamate at the cleft of the LBD bi-lobes are shown as green spheres. Schematic diagrams are shown to the right of each panel.

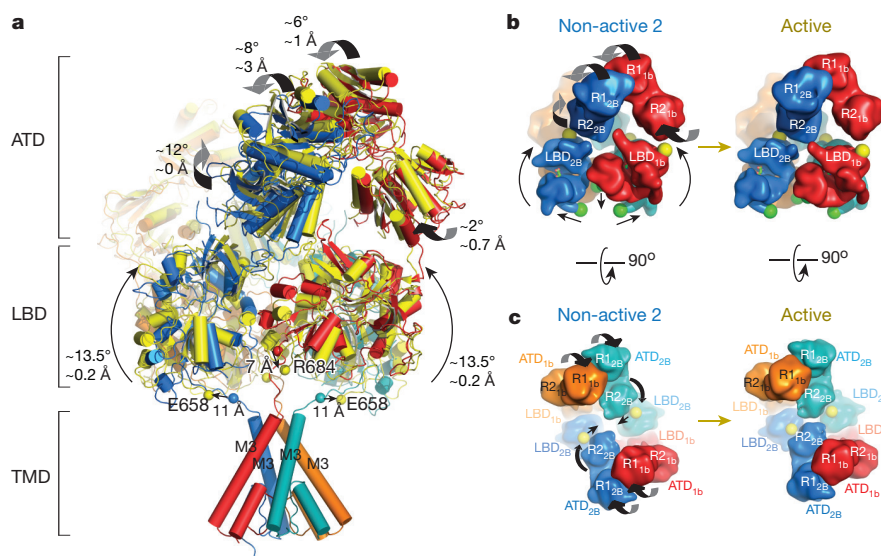


Figure 6 | Consorted movement of the ATD and LBD opens the gate. a, Structural comparison of NMDA receptors in the ‘non-active 2’ (same colour code as in Fig. 3) and the ‘active’ (yellow) conformations, as in Figs 4 and 5. The M1 and M4 helices of the TMD are omitted for clarity. The arrowed arcs indicate rotation from non-active 2 to active. The first ordered residues on the linker between the M3 helices on the TMD and

LBD in the active structure (GluN1b Arg684 and GluN2B Glu658) are shown as spheres. b, c, Schematic diagram viewed from the side of the tetramer (b) and top of the ATD (c). GluN1b Arg684 and GluN2B Glu658 are shown as green spheres and the residues at the ATD–LBD linker (GluN1b Ser416 and GluN2B His405) are shown as yellow spheres.

for the AMPA receptors¹⁷. This may indicate that the TMD domain is structurally more variable in activated receptors compared to non-activated receptors. Finally, the comparison of the cryo-EM classes with GluA2 AMPA receptor in the pre-open state, which represents a closed channel^{17,31}, shows that there is a greater difference between the active and pre-open states than between the non-active 2 and pre-open states (Extended Data Fig. 8), consistent with our observation that the TMD ion channel in the non-active structures are also closed.

Conclusion

We report conformational changes in multiple domains that are experimentally linked to activation of mammalian GluN1b–GluN2B NMDA receptors. The activation requires opening of the bi-lobed architecture of the GluN2B ATD and reorientation of the heterodimeric arrangement in the GluN1b–GluN2B ATD, as captured at high-resolution by the crystal structure presented here. These changes lead to rotated GluN1b–GluN2B heterodimeric pairs in both the ATD and LBD, causing dilation of the gating ring. The mechanistic understanding gained in the current study represents an important first step in understanding the sophisticated activation schemes^{26,33,34} that are essential for mammalian NMDA receptor function.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 14 December 2015; accepted 18 March 2016.

Published online 2 May 2016.

1. Traynelis, S. F. *et al.* Glutamate receptor ion channels: structure, regulation, and function. *Pharmacol. Rev.* **62**, 405–496 (2010).
2. Benveniste, M. & Mayer, M. L. Structure-activity analysis of binding kinetics for NMDA receptor competitive antagonists: the influence of conformational restriction. *Br. J. Pharmacol.* **104**, 207–221 (1991).
3. Clements, J. D. & Westbrook, G. L. Activation kinetics reveal the number of glutamate and glycine binding sites on the *N*-methyl-D-aspartate receptor. *Neuron* **7**, 605–613 (1991).
4. Johnson, J. W. & Ascher, P. Glycine potentiates the NMDA response in cultured mouse brain neurons. *Nature* **325**, 529–531 (1987).
5. Nowak, L., Bregestovski, P., Ascher, P., Herbet, A. & Prochiantz, A. Magnesium gates glutamate-activated channels in mouse central neurones. *Nature* **307**, 462–465 (1984).
6. Mayer, M. L., Westbrook, G. L. & Guthrie, P. B. Voltage-dependent block by Mg²⁺ of NMDA responses in spinal cord neurones. *Nature* **309**, 261–263 (1984).
7. Mayer, M. L., MacDermott, A. B., Westbrook, G. L., Smith, S. J. & Barker, J. L. Agonist- and voltage-gated calcium entry in cultured mouse spinal cord neurons under voltage clamp measured using arsenazo III. *J. Neurosci.* **7**, 3230–3244 (1987).
8. Granger, A. J. & Nicoll, R. A. Expression mechanisms underlying long-term potentiation: a postsynaptic view, 10 years on. *Phil. Trans. R. Soc. Lond. B* **369**, 20130136 (2014).
9. Paoletti, P., Bellone, C. & Zhou, Q. NMDA receptor subunit diversity: impact on receptor properties, synaptic plasticity and disease. *Nature Rev. Neurosci.* **14**, 383–400 (2013).
10. Armstrong, N., Sun, Y., Chen, G. Q. & Gouaux, E. Structure of a glutamate-receptor ligand-binding core in complex with kainate. *Nature* **395**, 913–917 (1998).
11. Furukawa, H., Singh, S. K., Mancusso, R. & Gouaux, E. Subunit arrangement and function in NMDA receptors. *Nature* **438**, 185–192 (2005).
12. Mayer, M. L. Crystal structures of the GluR5 and GluR6 ligand binding cores: molecular mechanisms underlying kainate receptor selectivity. *Neuron* **45**, 539–552 (2005).
13. Hansen, K. B., Furukawa, H. & Traynelis, S. F. Control of assembly and function of glutamate receptors by the amino-terminal domain. *Mol. Pharmacol.* **78**, 535–549 (2010).
14. Gielen, M., Siegler Retchless, B., Mony, L., Johnson, J. W. & Paoletti, P. Mechanism of differential control of NMDA receptor activity by NR2 subunits. *Nature* **459**, 703–707 (2009).
15. Yuan, H., Hansen, K. B., Vance, K. M., Ogden, K. K. & Traynelis, S. F. Control of NMDA receptor function by the NR2 subunit amino-terminal domain. *J. Neurosci.* **29**, 12045–12058 (2009).
16. Zhu, S. & Paoletti, P. Allosteric modulators of NMDA receptors: multiple sites and mechanisms. *Curr. Opin. Pharmacol.* **20**, 14–23 (2015).

17. Meyerson, J. R. *et al.* Structural mechanism of glutamate receptor activation and desensitization. *Nature* **514**, 328–334 (2014).
18. Karakas, E., Regan, M. C. & Furukawa, H. Emerging structural insights into the function of ionotropic glutamate receptors. *Trends Biochem. Sci.* **40**, 328–337 (2015).
19. Regan, M. C., Romero-Hernandez, A. & Furukawa, H. A structural biology perspective on NMDA receptor pharmacology and function. *Curr. Opin. Struct. Biol.* **33**, 68–75 (2015).
20. Sobolevsky, A. I., Rosconi, M. P. & Gouaux, E. X-ray structure, symmetry and mechanism of an AMPA-subtype glutamate receptor. *Nature* **462**, 745–756 (2009).
21. Kumar, J., Schuck, P. & Mayer, M. L. Structure and assembly mechanism for heteromeric kainate receptors. *Neuron* **71**, 319–331 (2011).
22. Karakas, E. & Furukawa, H. Crystal structure of a heterotetrameric NMDA receptor ion channel. *Science* **344**, 992–997 (2014).
23. Lee, C. H. *et al.* NMDA receptor structures reveal subunit arrangement and pore architecture. *Nature* **511**, 191–197 (2014).
24. Karakas, E., Simorowski, N. & Furukawa, H. Subunit arrangement and phenylethanolamine binding in GluN1/GluN2B NMDA receptors. *Nature* **475**, 249–253 (2011).
25. Karakas, E., Simorowski, N. & Furukawa, H. Structure of the zinc-bound amino-terminal domain of the NMDA receptor NR2B subunit. *EMBO J.* **28**, 3910–3920 (2009).
26. Banke, T. G. & Traynelis, S. F. Activation of NR1/NR2B NMDA receptors. *Nature Neurosci.* **6**, 144–152 (2003).
27. Farina, A. N. *et al.* Separation of domain contacts is required for heterotetrameric assembly of functional NMDA receptors. *J. Neurosci.* **31**, 3565–3579 (2011).
28. Jespersen, A., Tajima, N., Fernandez-Cuervo, G., Garnier-Amblard, E. C. & Furukawa, H. Structural insights into competitive antagonism in NMDA receptors. *Neuron* **81**, 366–378 (2014).
29. Vance, K. M., Simorowski, N., Traynelis, S. F. & Furukawa, H. Ligand-specific deactivation time course of GluN1/GluN2D NMDA receptors. *Nature Commun.* **2**, 294 (2011).
30. Sirrieh, R. E., MacLean, D. M. & Jayaraman, V. A. Conserved structural mechanism of NMDA receptor inhibition: A comparison of ifenprodil and zinc. *J. Gen. Physiol.* **146**, 173–181 (2015).
31. Dürr, K. L. *et al.* Structure and Dynamics of AMPA Receptor GluA2 in Resting, Pre-Open, and Desensitized States. *Cell* **158**, 778–792 (2014).
32. Kazi, R., Dai, J., Sweeney, C., Zhou, H. X. & Wollmuth, L. P. Mechanical coupling maintains the fidelity of NMDA receptor-mediated currents. *Nature Neurosci.* **17**, 914–922 (2014).
33. Popescu, G. & Auerbach, A. The NMDA receptor gating machine: lessons from single channels. *Neuroscientist* **10**, 192–198 (2004).
34. Popescu, G., Robert, A., Howe, J. R. & Auerbach, A. Reaction mechanism determines NMDA receptor response to repetitive stimulation. *Nature* **430**, 790–793 (2004).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank staff at the 23-ID beamlines at the Advanced Photon System in the Argonne National Laboratory. We are grateful to Z. Yu, C. Hong and R. Huang at the Janelia Research Center/HHMI EM facility for their support. This work was supported by the National Institutes of Health (MH085926 and GM105730), the Stanley Institute of Cognitive Genomics, Burroughs Wellcome Fund Collaborative Research Travel Grant, the Robertson Research Fund of Cold Spring Harbor Laboratory (all to H.F.), Japan Society for the Promotion of Science (to N.T.) and the Visiting Scientist program of the Janelia Research Center to allow H.F. to conduct cryo-EM work.

Author Contributions The authors jointly contributed to project design. N.T. and H.F. performed X-ray crystallography and electrophysiology. N.S. purified and characterized antibodies critical for the x-ray crystallographic study. T.G., R.D.A., N.G. and H.F. were involved in structural analysis by cryo-EM. E.K. expressed and purified proteins for the cryo-EM analysis and conducted model building and refinement of the cryo-EM structures. N.T., E.K., T.G., N.G. and H.F. were involved in manuscript preparation.

Author Information Atomic coordinates and structure factor for the apo-GluN1b–GluN2B ATD is deposited in the Protein Data Bank under the accession code 5B3J; the cryo-EM coordinates are deposited under the accession codes 5FXG, 5FXH, 5FXI, 5FXJ and 5FXK. The cryo-EM maps are deposited in EMDB under accession codes EMD-3352, EMD-3353, EMD-3354, EMD-3355 and EMD-3356. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to H.F. (furukawa@cshl.edu) or N.G. (grigorieffn@janelia.hhmi.org).

METHODS

Data reporting. No statistical methods were used to predetermine sample size. The experiments were not randomized and the investigators were not blinded to outcome assessment.

Production of GluN1b/GluN2B ATD, GluN1b/GluN2B NMDA receptors and Fab fragment. The constructs of GluN1b and GluN2B ATDs are identical to those used in our previous study, and were expressed and purified in the same way²⁴. The purified protein was deglycosylated by endoglycosidase F1. Monoclonal antibodies (mouse IgG) that bind rat GluN2B ATD were obtained by immunizing mice with the purified intact GluN1–GluN2B NMDA receptors using the standard protocol. IgGs were purified from hybridoma cell culture supernatant by Protein-A Sepharose (GE healthcare). Fab fragments of the antibody were obtained by papain proteolysis followed by re-chromatographing onto Protein-A Sepharose to remove the Fc region. The purified GluN1b–GluN2B ATD and Fab were mixed, and the ATD–Fab complex was isolated by Superdex200 (10/300; GE Healthcare). The intact tetrameric GluN1b–GluN2B NMDA receptors were expressed and purified as previously described²².

Crystallization, data collection and structural determination of apo-GluN1b–GluN2B ATD. The purified GluN1b–GluN2B ATD–Fab complex was concentrated to 8 mg ml^{−1} and dialysed against a buffer containing 10 mM Tris–HCl (pH 8.0) and 50 mM NaCl. The crystals were grown at 18 °C by the hanging-drop vapour diffusion method. GluN1b–GluN2B ATD–Fab complex was mixed with a half volume of reservoir solution (3–5 µl total drop size), which contained 0.1 M sodium acetate (pH 4.5), 27% PEG3350, 2.2 M sodium formate, and 0.05 M calcium chloride. Cryoprotection was achieved by supplying 8% glycerol to the crystallization condition. Crystals were flash-frozen in liquid nitrogen. Data sets were collected at the wavelength of 1.0 Å and at the 23ID-D beamline in the Advanced Photon System in the Argonne National Laboratory and processed using HKL2000 (ref. 35) (Extended Data Table 1). The crystal structure of GluN1b–GluN2B ATD–Fab17 complex was solved by molecular replacement using the coordinate of GluN1b/GluN2B ATD (PDB ID, 3QEL) and Fab (PDB ID, 1BAF) and by using the program Phaser³⁶. The model refinement was performed using the program Phenix³⁷.

Electrophysiology. GluN1–4b/GluN2B NMDA receptors were expressed by injecting cRNAs at a 1:2 ratio (GluN1:GluN2, w/w) into defolliculated *Xenopus laevis* oocytes (0.05–0.15 ng total per oocyte). After 24–48 h incubation at 18 °C, currents were measured by two-electrode voltage clamp in a solution containing 5 mM HEPES, 100 mM NaCl, 0.3 mM BaCl₂ and 10 mM Tricine at pH 6.5 (adjusted with KOH) using agarose-tipped microelectrode (0.4–1.0 MΩ) at the holding potential of −60 mV. Currents were evoked by application of 100 µM glycine and L-glutamate. For MTS experiments, fresh stock of MTS reagents were made and added to the recording buffers at the final concentration of 200 µM. The data were analysed using the program Pulse (HEKA) and the graphs were generated by the program Kaleidagraph (Synergy).

Cysteine crosslinking and western blot. Recombinant wild-type and mutant GluN1–4b/GluN2B NMDA receptors (GluN2B CTD truncated as in Extended Data Fig. 1), were expressed in the *Spodoptera frugiperda* (Sf9) baculovirus system as described previously²². The infected cell pellets were solubilized in a buffer containing 50 mM HEPES pH 7.3, 200 mM NaCl, 0.5% LMN, and 1 mM PMSE. The GluN1–4b/GluN2B NMDA receptor proteins were purified by Strep-Tactin Sepharose (IBA) and subjected to 7% SDS-polyacrylamide gel electrophoresis in the presence and absence of 100 mM β-mercaptoethanol. The proteins were transferred to nitrocellulose membranes (GE healthcare). The membranes were blocked with 5% milk in a phosphate saline buffer containing 0.05% Tween-20, incubated with mouse monoclonal anti-GluN1 antibody (MAB1586, Millipore) or anti-GluN2B antibody (AB93610, Abcam), followed by horseradish peroxidase (HRP)-conjugated anti-mouse secondary antibodies (GE healthcare). The ECL detection kit (GE healthcare) was used to visualize bands.

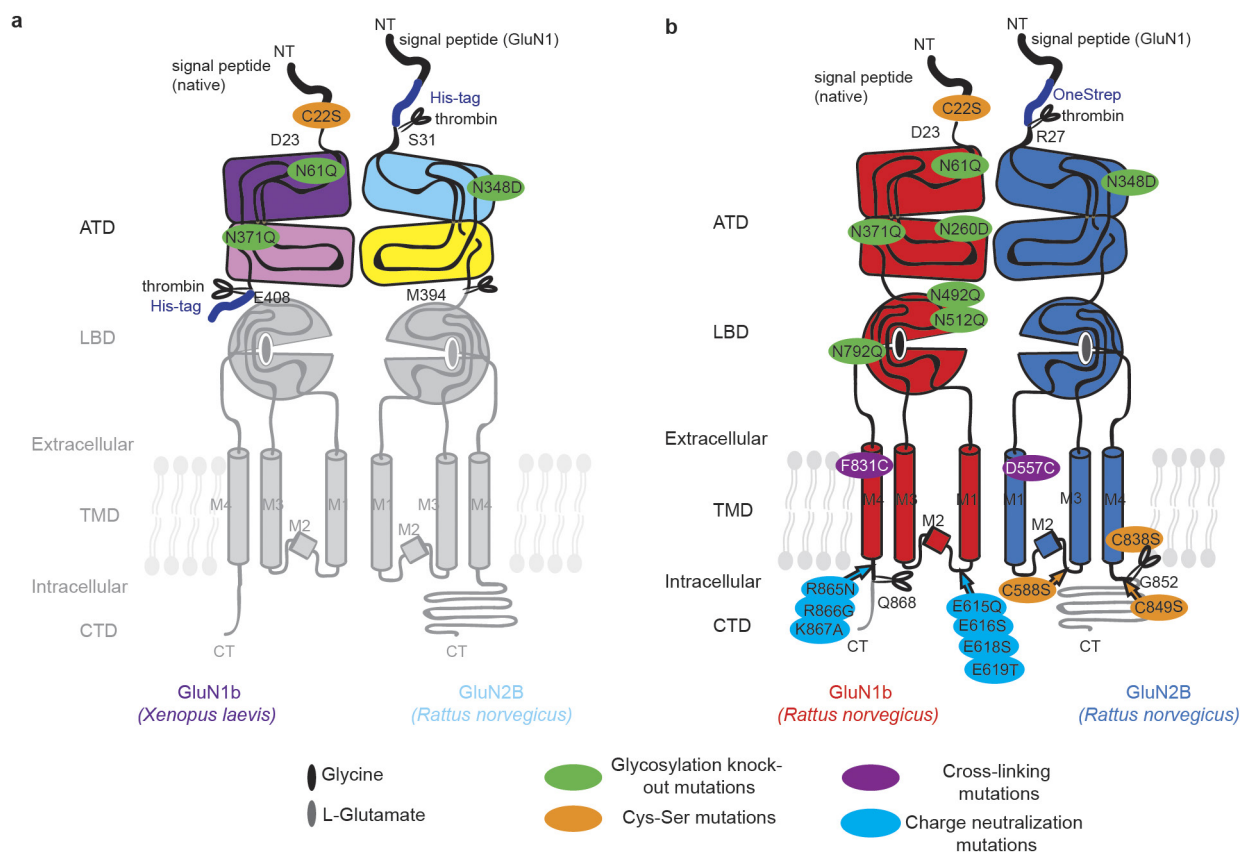
Cryo-EM specimen preparation and image acquisition. Purified GluN1b/GluN2B NMDA receptor at 2 mg ml^{−1} was placed on C-flat 1.2/1.3 Cu 400 mesh

grids (Protochips), which had previously been subjected to glow discharge for 45 s at 15 mA, and plunge-frozen using an FEI Vitrobot Mark 2 with a 3 s blot time and at relative humidity between 85% and 95%. The data were collected on an FEI Titan Krios microscope operating at 300 kV. 1,204 movies were collected on a Gatan K2 Summit direct electron detector (Gatan, Inc.) in super resolution mode with a pixel size of 0.655 Å per super resolution pixel. Each exposure was 21 s long and recorded as a movie of 70 frames. The exposure per frame as reported by Digital Micrograph (Gatan, Inc.) was ~1.4 e[−] per Å², which corresponds to an exposure of ~8 electrons per pixel per second on the camera. Videos were recorded at a range of underfocus between ~1.0 µm and ~2.5 µm.

Image processing. Super-resolution movie frames were initially corrected for magnification distortion³⁸. The frames were then downsampled by a factor of 2 using Fourier cropping to a pixel size of 1.31 Å, motion-corrected and exposure filtered using Unblur³⁹ and the microscope CTF was determined using CTFFIND4⁴⁰ on motion-corrected but non-exposure filtered movie sums. Around 90,000 particles were picked automatically then verified manually from the aligned movie sums which had been exposure filtered, but not noise restored, resulting in a strong low-pass filter. The picked particles were extracted into 256 × 256 boxes. Initial particle alignment parameters were assigned by a brute force search in FREALIGN v9⁴¹, sampling every 5° and limiting the resolution to 15 Å using a previously determined structure as a reference. These parameters were further refined and classified into six 3D classes with FREALIGN. For classes 1, 3 and 6, the highest resolution included in the alignment was 8 Å, for class 4 the highest included resolution was 12 Å, and for class 5 it was 6.5 Å. Class 2 showed only low-resolution features and was discarded. The resulting resolutions as determined by the 0.143 cut-off⁴² were 5.0–6.7 Å (Extended Data Fig. 6). Maps were rendered using UCSF Chimera⁴³, after applying a bfactor of −600 Å².

Model building. The GluN1a–GluN2B crystal structure (PDB ID, 4PE5)²² was docked into the cryo-EM maps followed by rigid-body fitting of the individual ATD R1 and R2 lobes and LBDs of both GluN1 and GluN2B into the cryo-EM maps using Coot⁴⁴. Both the rat GluN1a–GluN2B crystal structure (PDB ID, 4PE5)²² and *Xenopus* GluN1–GluN2B NMDA receptor (PDB ID, 4TLM)²³ were used to model the TMD. The resulting models were manually modified to fit into the density using Coot⁴⁴ and refined against the cryo-EM maps using Phenix real space refinement⁴⁵. Refinement statistics are shown in Extended Data Table 2. Class X and class Y are similar to 'non-active 2'.

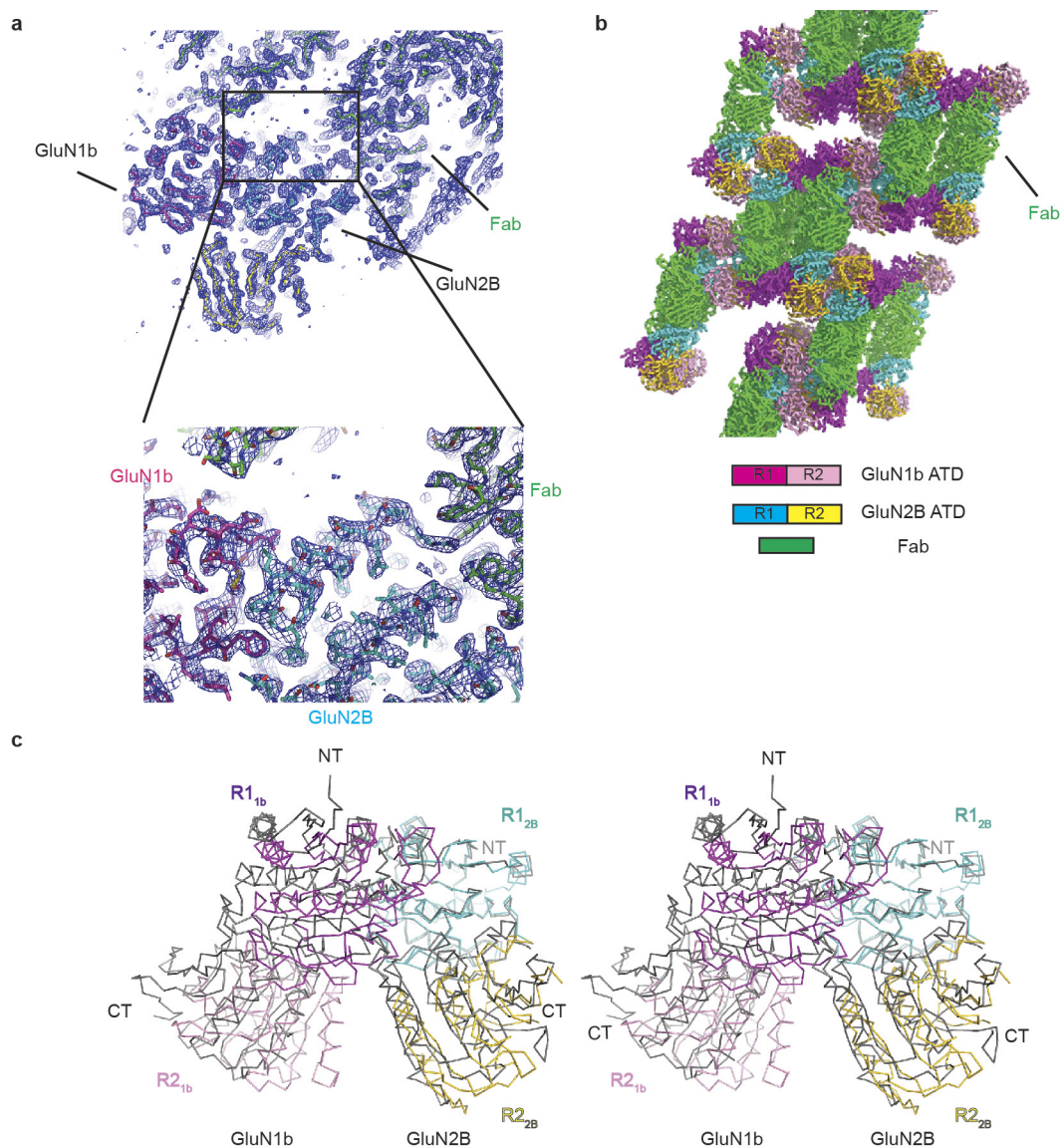
35. Otwinowski, Z. & Minor, W. Processing of X-ray diffraction data collected in oscillation mode. *Methods Enzymol.* **276**, 307–326 (1997).
36. McCoy, A. J. *et al.* Phaser crystallographic software. *J. Appl. Cryst.* **40**, 658–674 (2007).
37. Adams, P. D. *et al.* PHENIX: building new software for automated crystallographic structure determination. *Acta Crystallogr. D* **58**, 1948–1954 (2002).
38. Grant, T. & Grigorieff, N. Automatic estimation and correction of anisotropic magnification distortion in electron microscopes. *J. Struct. Biol.* **192**, 204–208 (2015).
39. Grant, T. & Grigorieff, N. Measuring the optimal exposure for single particle cryo-EM using a 2.6 Å reconstruction of rotavirus VP6. *eLife* **4**, e06980 (2015).
40. Rohou, A. & Grigorieff, N. CTFFIND4: fast and accurate defocus estimation from electron micrographs. *J. Struct. Biol.* **192**, 216–221 (2015).
41. Lyumkis, D., Brilot, A. F., Theobald, D. L. & Grigorieff, N. Likelihood-based classification of cryo-EM images using FREALIGN. *J. Struct. Biol.* **183**, 377–388 (2013).
42. Rosenthal, P. B. & Henderson, R. Optimal determination of particle orientation, absolute hand, and contrast loss in single-particle electron cryomicroscopy. *J. Mol. Biol.* **333**, 721–745 (2003).
43. Pettersen, E. F. *et al.* UCSF Chimera—a visualization system for exploratory research and analysis. *J. Comput. Chem.* **25**, 1605–1612 (2004).
44. Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. Features and development of Coot. *Acta Crystallogr. D* **66**, 486–501 (2010).
45. Afonine, P. V., Headd, J. J., Terwilliger, T. C. & Adams, P. D. New tool: phenix.real_space_refine. *Computational Crystallography Newsletter* **4**, 43–44 (2013).



Extended Data Figure 1 | Domain organization and constructs.

a, The construct design for GluN1b and GluN2B ATD used in this study. GluN1b from *Xenopus laevis* is combined with GluN2B from rat, as in

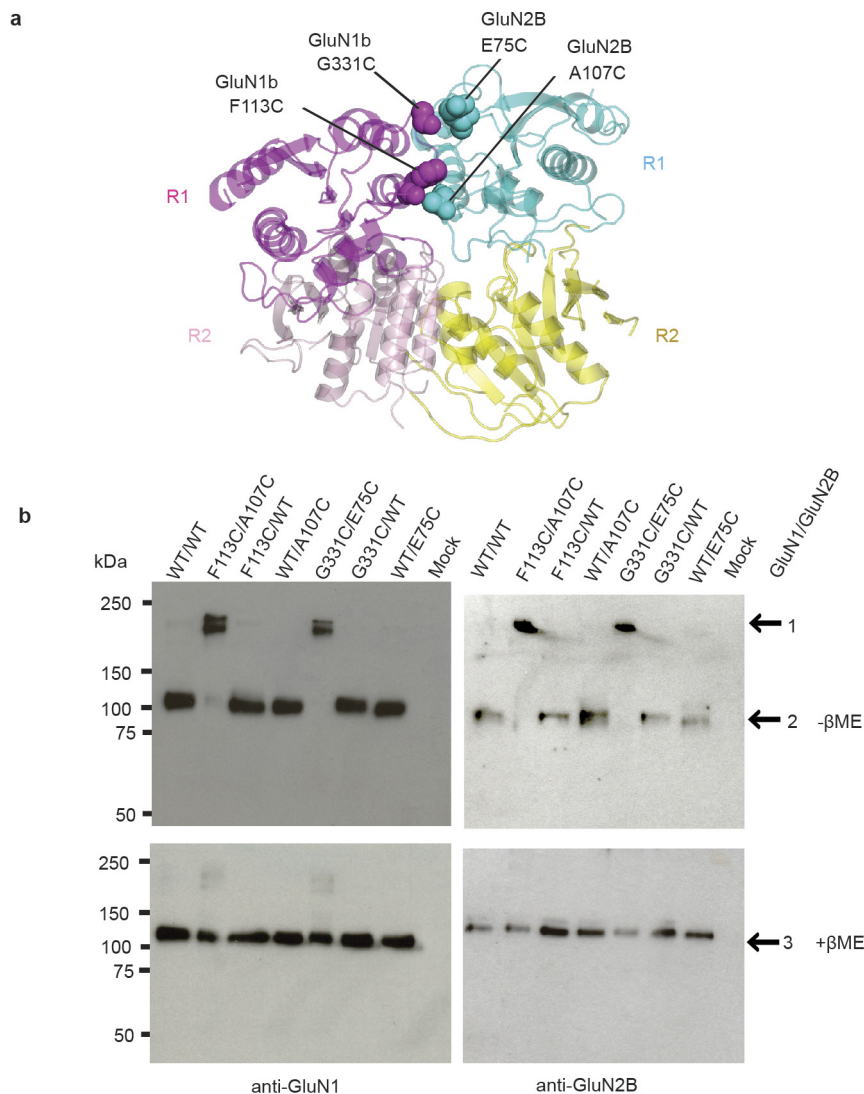
the previous study on the ATD²⁴. **b**, The construct design for the intact GluN1b/GluN2B NMDA receptors from rat. A similar construct was used in previous studies and shown to be fully functional²².



Extended Data Figure 2 | Structure of the apo-GluN1b-GluN2B ATD.

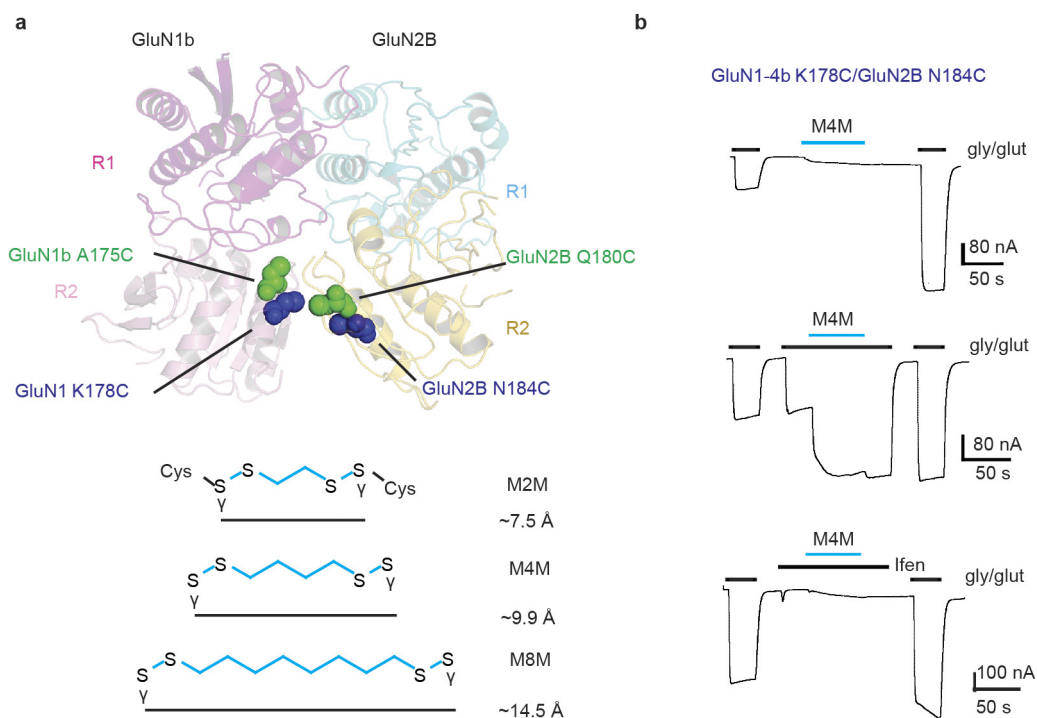
a, Representative $2F_o - F_c$ electron density map contoured at 1.2σ showing continuous density throughout GluN1b, GluN2B and Fab. The quality of the electron density map is at a sufficient level to model amino acid side chains (see lower panel). **b**, Crystal packing of GluN1b-GluN2B ATD-Fab showing that the packing is mediated robustly by Fab molecules (green).

The colour coding for the ATD is the same as in Fig. 1. **c**, Comparison of the apo-GluN1b-GluN2B ATD and ifenprodil-GluN1b-GluN2B ATD (grey) by stereo presentation. Colour coding for the apo-GluN1b-GluN2B ATD is the same as in Fig. 1. Here the two structures are superimposed at GluN2B R1.



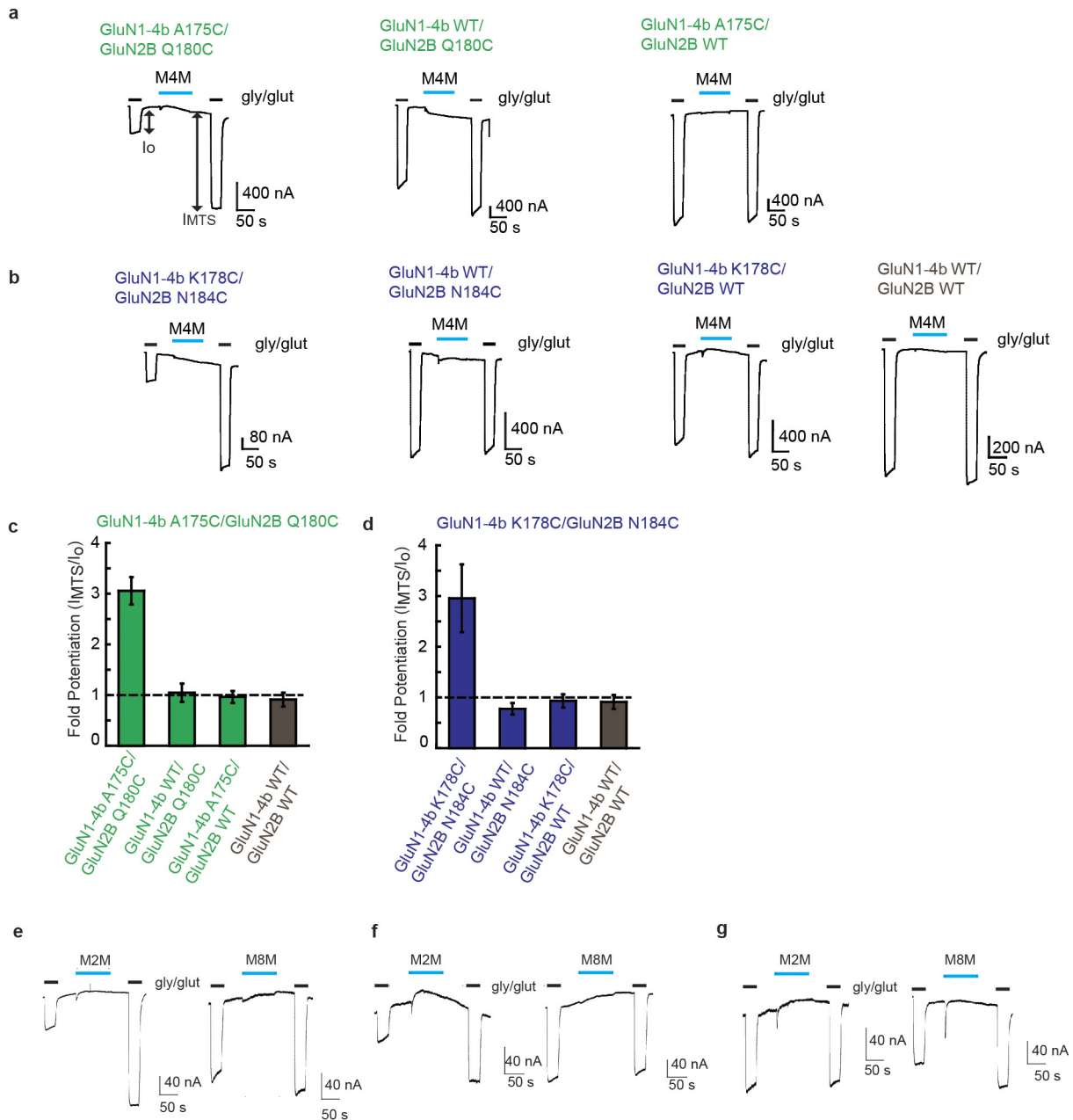
Extended Data Figure 3 | Validation of the crystal structure by disulfide cross-linking. **a**, Crystal structure of the apo-GluN1b–GluN2B ATD showing locations of the mutated residues, GluN1b Phe113, GluN1b Gly331, GluN2B Ala107 and GluN2B Glu75 in spheres. **b**, Western blots using anti-GluN1 (left) and anti-GluN2B (right) antibodies on purified intact GluN1b/GluN2B NMDA receptor that lacks the CTD.

Upper and lower panels are blots run in the absence and presence of β -mercaptoethanol (β ME), respectively. Bands highlighted by arrow 1 are consistent with the molecular weight of GluN1–GluN2B heterodimers, whereas those highlighted by arrows 2 and 3 are consistent with the molecular weights of monomers of GluN1–4b and GluN2B.



Extended Data Figure 4 | Conformational trap shows the apo-GluN1b-GluN2B ATD structure to represent ‘active’ form-II. **a**, Location of engineered cysteines in the crystal structure of the apo-GluN1b-GluN2B ATD. The cysteine mutant pairs, GluN1-4b(Ala175Cys)/GluN2B(Gln180Cys) (green spheres) and GluN1-4b(Lys178Cys)/GluN2B(Asn184Cys) (blue spheres) are co-expressed in *Xenopus* oocytes and cross-linked by bifunctional MTS with different linker lengths (M2M, M4M and M8M).

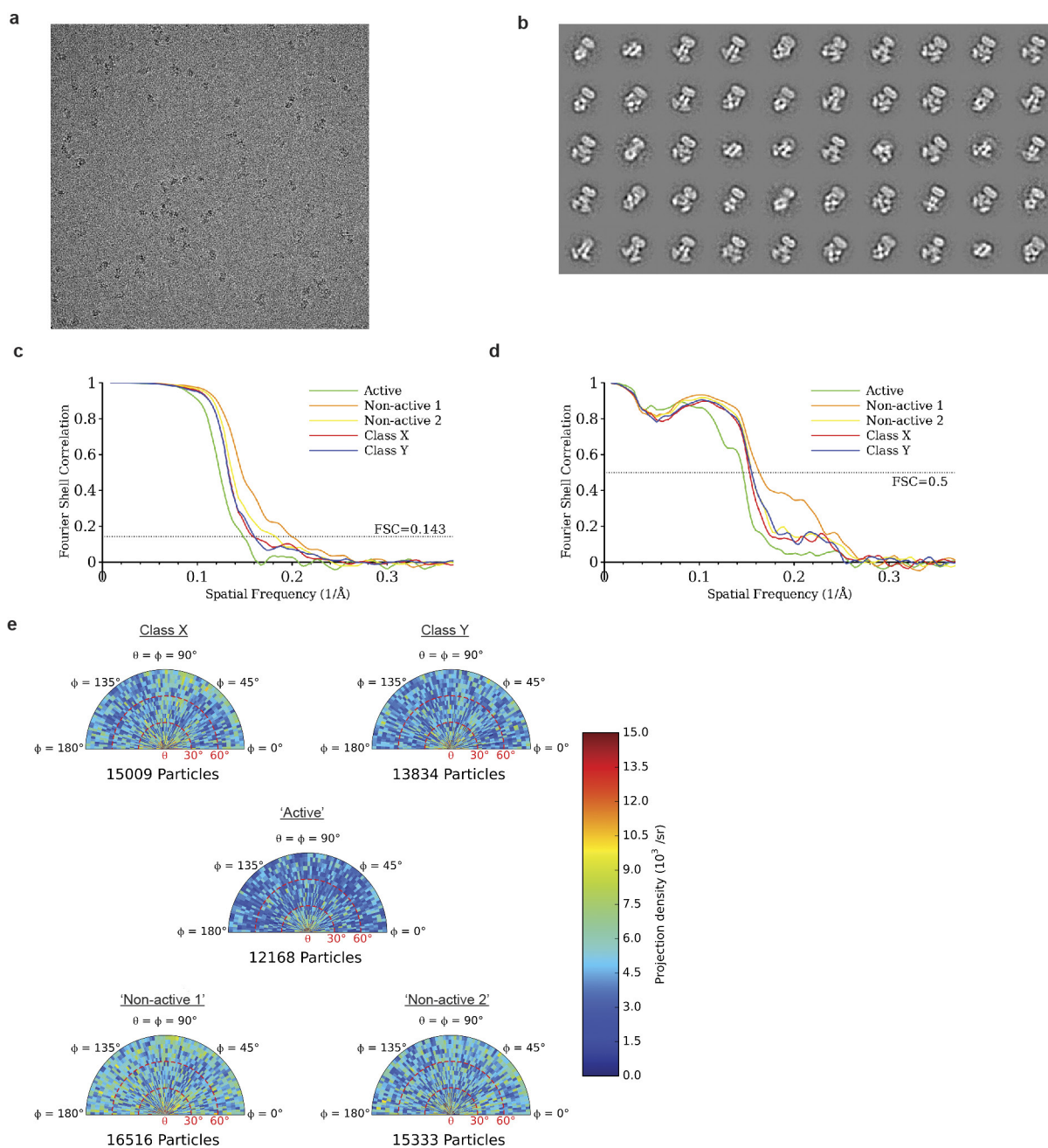
b, Application of 200 μ M M4M in the presence or absence of 100 μ M agonists (glycine (gly)/glutamate (glut)) potentiates the macroscopic current measured at the holding potential of -60 mV by two-electrode voltage clamp. No potentiation was observed when M4M was applied in the presence of ifenprodil (ifen). Shown here are the representative recording profiles for the GluN1-4b(Lys178Cys)/GluN2B(Asn184Cys) pair.



Extended Data Figure 5 | Effect of bi-MTAs on cysteine mutants.

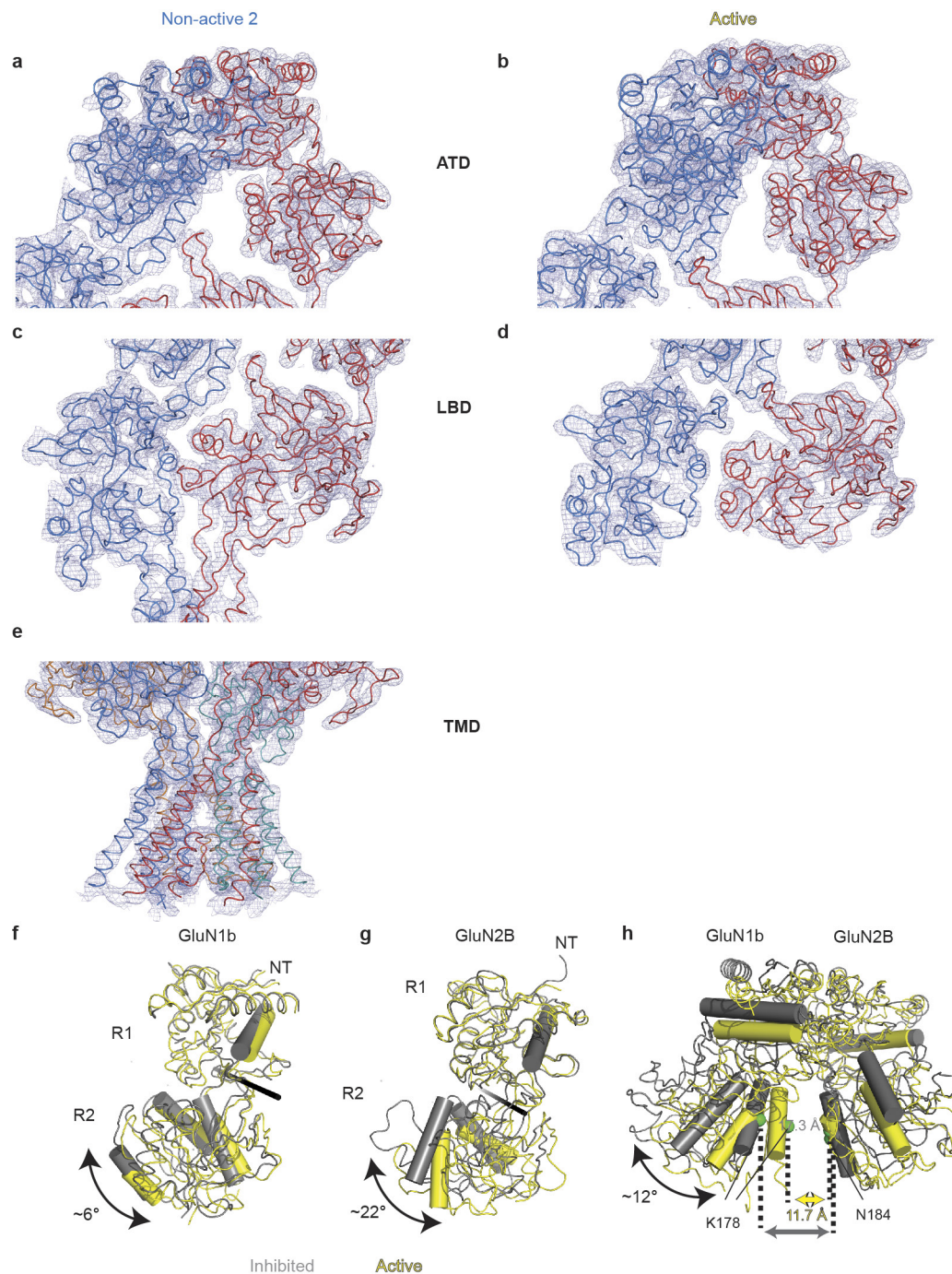
a, b, M4M specifically traps the active conformation at the engineered cysteines. Representative electrophysiological traces for the mutant pairs, GluN1-4b(Ala175Cys)/GluN2B(Gln180Cys) (green spheres) and GluN1-4b(Lys178Cys)/GluN2B(Asn184Cys) as well as mutant and wild-type pairs. The experiments were conducted by two-electrode voltage clamp as in Fig. 2. The potentiation by M4M (represented by I_{MTS}/I_o) is only observed when both GluN1 and GluN2B cysteine mutants are co-expressed. No potentiation was observed when the cysteine mutant of one subunit is combined with the wild type of the other, indicating that the

effect of M4M modification is specific and validating the relevance of the experiments. **c, d**, Bar graphs presenting the degree of potentiation from the recordings in **a** and **b**. Error bars represent \pm s.d. for data obtained from five different oocytes per mutant combination. **e–g**, M2M but not M8M potentiates the mutant GluN1b–GluN2B NMDA receptor. The same experiment as above or in Fig. 2 was conducted using M2M or M8M on GluN1-4b(Ala175Cys)/GluN2B(Gln180Cys) (**e**) GluN1-4b(Lys178Cys)/GluN2B(Asn184Cys) (**f**), and wild-type GluN1-4b/GluN2B (**g**). Shown are representative electrophysiological recordings used to estimate the degree of bi-MTS potentiation presented in Fig. 2c.



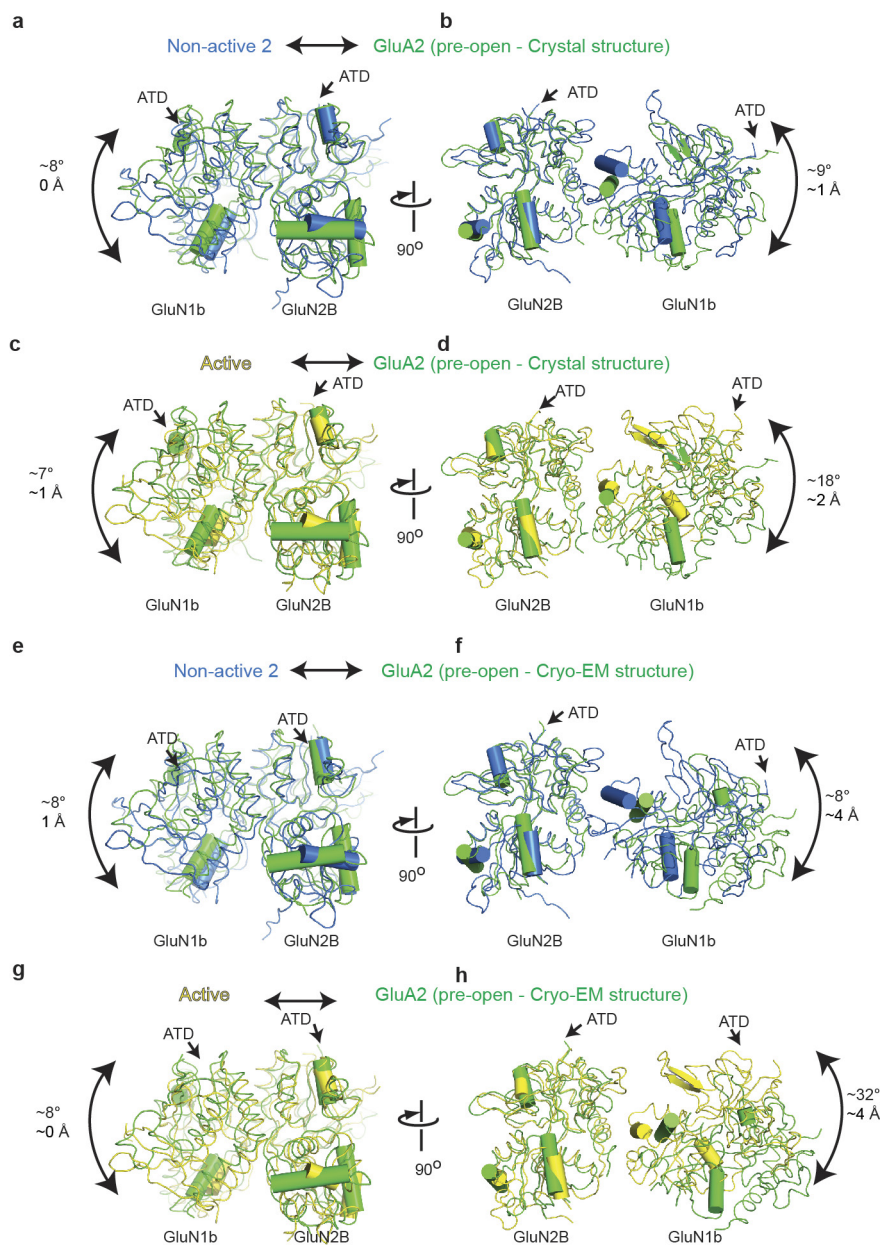
Extended Data Figure 6 | Cryo-EM analysis on GluN1b/GluN2B NMDA receptors. **a**, Representative motion-corrected image collected at 22,500 magnification. **b**, Two-dimensional class averages. **c**, **d**, Fourier shell correlation curves for unmasked data (**c**) and model versus electron microscopy map (**d**). Class X and Y are similar to 'non-active 2'

and were not further analysed. **e**, Orientation plots for each class, plotting the distribution of Euler angles assigned to all particles contributing to that class with an occupancy of at least 80%. For each class, the number of particles which have that class as their highest occupancy value is also shown.



Extended Data Figure 7 | Representative cryo-EM density, model fit and structural comparison of the ATD in inhibited and active conformations of cryo-EM structures. a–e, Here, the cryo-EM maps for non-active 2 and active 3D classes are shown along with the refined models. Densities are shown at the ATD and LBD (a–d) for both of the 3D classes and at the TMD (e) for non-active 2. **f, g,** Superimposition of R1 lobes of GluN1b (f) and GluN2B (g) illustrates the relative ‘opening’ between R1 and R2 lobes in the inhibited and active forms of intact NMDA receptors. The extent of GluN2B ATD opening is similar to that observed between the crystal structures of the ifenprodil–GluN1b–GluN2B ATD and the apo–GluN1b–GluN2B ATD as in Fig. 1. GluN1b and

GluN2B ATDs are shown in grey and yellow for the inhibited and active states, respectively. **h,** Comparison of the GluN1b–GluN2B ATD heterodimers between ifenprodil inhibited and active cryo-EM structures. Superimposition of the GluN2B R1 lobes reveals an $\sim 12^\circ$ rotation of the GluN1b ATD relative to the GluN2B ATD in the similar manner to the crystal structure of the apo–GluN1b–GluN2B ATD as in Fig. 1. The black rods indicate the axis of rotation between the two cryo-EM structures. The distance of the R2 lobes in the GluN1b–GluN2B heterodimers is measured between C α atoms of GluN1b(Lys178) and GluN2B(Asn184) (green spheres).



Extended Data Figure 8 | Structural comparison of the GluN1–GluN2B LBD in non-active and active conformations to the GluA2 LBD in pre-open state. **a–d**, The crystal structure of GluA2 AMPA receptor in the pre-open state (PDB ID, 4U1W; shown in green) aligned with the structures of GluN1–GluN2B in the non-active 2 (blue) (**a**, **b**) and active conformation (yellow) (**c**, **d**) by superimposing the LBDs of GluN2B onto GluA2. **e–h**, The equivalent superimposition with the cryo-EM structure of GluA2 AMPA receptor in the pre-open state (PDB ID, 4UQ6; shown in green). The overlaid structures are viewed through the LBD heterodimer interface (**a**, **c** or **e**, **g**) and the dimer of heterodimer interface (**b**, **d** or **f**, **h**). Here, the GluN2B LBD of the GluN1b–GluN2B NMDA receptor is superimposed onto the LBD of the GluA2 AMPA receptor and the shift of

the GluN1 LBD is measured with respect to the other GluA2 LBD. The homodimeric arrangement of GluA2 AMPA receptor in the pre-open state is similar to the heterodimeric arrangements of GluN1b–GluN2B NMDA receptors in both non-active 2 and active states (**a**, **c** or **e**, **g**). However, when the dimer of homodimer arrangement of GluA2 AMPA receptor is compared to the dimer of heterodimers arrangement of the GluN1b–GluN2B NMDA receptor, a greater difference is observed for the active NMDA receptor (**d**, **h**) than for the non-active 2 NMDA receptor (**b**, **f**). Here, the non-active 2 NMDA receptor as in Fig. 3 is subjected to superimposition. The non-active 1 and non-active 2 NMDA receptors have similar subunit arrangements in the LBD. The numbers in each panel represent degrees of rotations and translations.

Extended Data Table 1 | Data collection and refinement statistics for X-ray crystallography

Apo- GluN1b/GluN2B ATD – Fab17	
Data collection	
Space group	C2
Cell dimensions	
<i>a</i> , <i>b</i> , <i>c</i> (Å)	247.4, 80.0, 181.4
α , β , γ (°)	90.0, 127.2, 90.0
Resolution (Å)	50-2.90(2.93-2.90) *
<i>R</i> _{merge}	0.099 (0.602)
<i>I</i> / σ <i>I</i>	8.5 (1.84)
Completeness (%)	91.4 (93.0)
Redundancy	4.0 (3.5)
Refinement	
Resolution (Å)	30-2.9
No. reflections	52,910
<i>R</i> _{work} / <i>R</i> _{free}	0.273/0.302
No. atoms	
Protein	15,899
Ion (Na)	1
Water	112
B-factors	
Protein	46.3
Ligand/ion	52.1
Water	37.9
R.m.s deviations	
Bond lengths (Å)	0.008
Bond angles (°)	1.068

*Highest resolution shell is shown in parentheses.

Extended Data Table 2 | Refinement statistics for the cryo-EM structures

	Active	Non-Active 1	Non-Active 2	Class X	Class Y
PDB ID	5FXG	5FXH	5FXI	5FXJ	5FXK
EMDB ID	EMD-3352	EMD-3353	EMD-3354	EMD-3355	EMD-3356
Refinement					
Resolution (Å)	6.8	6.1	6.4	6.5	6.4
Map to Model CC	0.77	0.79	0.80	0.78	0.80
No. atoms					
Protein	12,693	15,578	15,381	15,475	15,598
R.m.s deviations					
Bond lengths (Å)	0.003	0.006	0.003	0.004	0.005
Bond angles (°)	0.55	0.55	0.60	0.58	0.55
Ramachandran					
Favored (%)	88.3	90.4	90.2	88.0	88.8
Allowed (%)	11.5	9.2	9.4	11.4	10.7
Disallowed (%)	0.2	0.5	0.4	0.6	0.5

Structure of spinach photosystem II–LHCII supercomplex at 3.2 Å resolution

Xuepeng Wei^{1,2*}, Xiaodong Su^{1*}, Peng Cao¹, Xiuying Liu^{1,2}, Wenrui Chang¹, Mei Li¹, Xinzhen Zhang¹ & Zhenfeng Liu¹

During photosynthesis, the plant photosystem II core complex receives excitation energy from the peripheral light-harvesting complex II (LHCII). The pathways along which excitation energy is transferred between them, and their assembly mechanisms, remain to be deciphered through high-resolution structural studies. Here we report the structure of a 1.1-megadalton spinach photosystem II–LHCII supercomplex solved at 3.2 Å resolution through single-particle cryo-electron microscopy. The structure reveals a homodimeric supramolecular system in which each monomer contains 25 protein subunits, 105 chlorophylls, 28 carotenoids and other cofactors. Three extrinsic subunits (PsbO, PsbP and PsbQ), which are essential for optimal oxygen-evolving activity of photosystem II, form a triangular crown that shields the Mn_4CaO_5 -binding domains of CP43 and D1. One major trimeric and two minor monomeric LHCII associate with each core-complex monomer, and the antenna–core interactions are reinforced by three small intrinsic subunits (PsbW, PsbH and PsbZ). By analysing the closely connected interfacial chlorophylls, we have obtained detailed insights into the energy-transfer pathways between the antenna and core complexes.

Powered by solar energy, plants, algae and cyanobacteria convert water and carbon dioxide into organic matter and release oxygen through photosynthesis. In oxygenic photosynthesis, the initial photophysical and photochemical processes are primarily mediated by two photosystems: photosystems I (PSI) and II (PSII)¹. PSII is a supramolecular complex embedded within the thylakoid membrane. It contains numerous protein subunits and various cofactors, including chlorophylls, carotenoids, an Mn_4CaO_5 cluster, a haem, plastoquinones and lipids². A characteristic functional feature of PSII is its ability to extract electrons from water molecules through a light-induced water-oxidizing reaction catalysed by the Mn_4CaO_5 cluster³. To collect photon energy and drive the photochemical reactions, plant PSII contains a series of peripheral light-harvesting complexes (the major light-harvesting complexes of PSII (LHCII) and minor ones named chlorophyll-binding protein of 29, 26 and 24 kDa (CP29, CP26 and CP24))^{4,5}. These antenna complexes surround the core complex of PSII, absorb light energy and transmit it to the reaction centre to induce charge separation in the special pair of chlorophylls named P680 (ref. 6).

The highest resolution at which the crystal structure of cyanobacterial PSII has been solved is 1.9 Å (refs 7, 8), after a decade-long optimization process starting from 3.8 Å resolution (reviewed in ref. 2). The structures of cyanobacterial PSII provide solid foundations for understanding the pathways of excitation energy transfer, electron transport and water splitting processes occurring within the complex. Although plant PSII has a core complex similar to that of cyanobacterial PSII, there are major differences in their luminal extrinsic domains and peripheral antenna systems. The structures of partial and intact core complexes of plant PSII have been solved through cryo-electron crystallography at 8–10 Å resolution^{9–11}, and X-ray structures of isolated LHCII and CP29 are available at 2.5–2.8 Å resolution^{12–14}. Furthermore, a 3D map of the PSII–LHCII supercomplex at 17 Å resolution was obtained through single-particle cryo-electron microscopy (cryo-EM)¹⁵, and 2D projection maps of larger supercomplexes with more antenna complexes bound have been reported at 12–13 Å (refs 16, 17). Nevertheless, the precise pathways of excitation energy

transfer between the peripheral antennae and core complex of plant PSII remain largely unclear owing to the lack of a high-resolution structure of the PSII–LHCII supercomplex. Moreover, the structural roles and mutual interactions of three important major extrinsic subunits (PsbO, PsbP and PsbQ) in plant PSII are unknown. Here we present a 3.2 Å resolution cryo-EM structure of spinach PSII in complex with LHCII, CP29, CP26 and four extrinsic proteins. Unprecedented details concerning the specific interactions between different components within the supercomplex are revealed.

Overall architecture

The PSII–LHCII supercomplex sample for the cryo-EM study was purified from spinach leaves. Its spectroscopic features, protein composition and pigment content analysis results are summarized in Extended Data Fig. 1. From a total of 1,774 cryo-EM micrographs collected (Fig. 1a), 192,071 particles were picked for further data processing. After 2D and 3D classification (Fig. 1b), 109,042 C₂S₂-type (C: PSII core complex; S: strongly associated LHCII trimer) particles were selected and processed with local motion correction and image polishing, leading to a cryo-EM map at an overall resolution of 3.2 Å (Fig. 1c, d and Extended Data Fig. 2; see Methods for more details). The actual resolution within the C₂S₂-type PSII–LHCII supercomplex varies from 3.0 Å in the core region to 4.0 Å in some of the peripheral regions (Extended Data Fig. 2b). An atomic model of the PSII–LHCII supercomplex has been constructed and refined against the 3.2 Å cryo-EM map (Extended Data Fig. 2c; see Methods for details).

As shown in Fig. 2, the spinach C₂S₂-type PSII–LHCII supercomplex forms a homodimer with two-fold symmetry running through the centre along the membrane normal. The dimerization interface in the core region closely resembles those of cyanobacterial PSII^{7,18,19}, and LHCII and CP29 at the peripheral regions extend the interface and enhance the stability of the dimeric supercomplex. Each monomer contains a core complex composed of four large intrinsic subunits (D1, D2, CP43 and CP47), twelve low-molecular-mass membrane-spanning subunits (PsbE, PsbF, PsbH, PsbI, PsbJ, PsbK, PsbL, PsbM,

¹National Laboratory of Biomacromolecules, CAS Center for Excellence in Biomacromolecules, Institute of Biophysics, Chinese Academy of Sciences, Beijing 100101, China. ²University of Chinese Academy of Sciences, Beijing 100049, China.

*These authors contributed equally to this work.

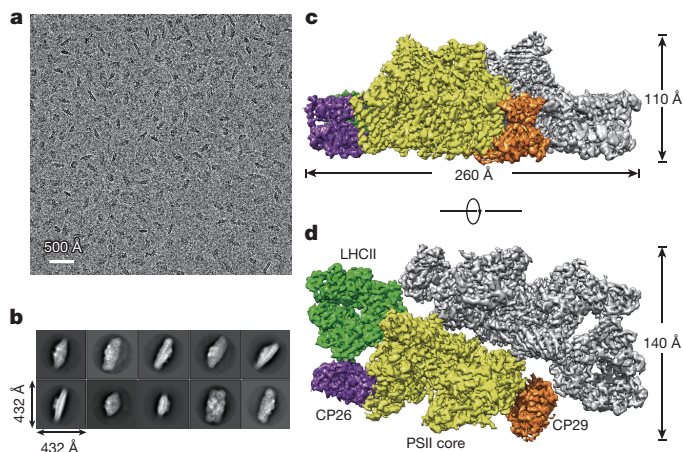


Figure 1 | Single-particle cryo-EM analyses of spinach PSII-LHCII supercomplex. **a**, A representative cryo-EM image of a spinach PSII-LHCII supercomplex sample. **b**, Selected 2D class averages of C_2S_2 -type PSII-LHCII supercomplex particles. **c**, **d**, 3D cryo-EM density map of the PSII-LHCII supercomplex. **c**, Side view along membrane normal. **d**, Bottom view from luminal side along membrane normal. One monomer of the supercomplex is shown in a four-colour mode with the PSII core in yellow, LHCII in green, CP26 in purple and CP29 in orange; the other monomer is shown in a single grey colour.

PsbTc, PsbW, PsbX and PsbZ) (Fig. 2a–c), and four extrinsic subunits attached on the luminal surface (PsbO, PsbP, PsbQ and PsbTn) (Fig. 2d). The major intrinsic subunits (D1, D2, CP43 and CP47) of spinach PSII core share high similarities with those of cyanobacterial PSII. The amino acid sequences of spinach D1 (PsbA), D2 (PsbD), CP47 (PsbB) and CP43 (PsbC) proteins are 85%, 90%, 77% and 84%, respectively, identical to those from *Thermosynechococcus vulcanus* photosystem II (TvPSII) (ref. 7). Their structures can be superposed on those of TvPSII with root mean square deviations (r.m.s.d.) of α -carbon atoms at 0.57 (D1), 0.64 (D2), 0.68 (CP47) and 0.69 Å (CP43), respectively (Extended Data Fig. 3a). Moreover, the cofactor binding sites within the core complex are well conserved between the two species.

Surrounding the four major core subunits, twelve low-molecular-mass intrinsic subunits form a discontinuous belt-like structure wrapping around the core (Fig. 2a and Extended Data Fig. 3b, c). Among them, eleven can find their homologues on similar binding sites in cyanobacterial PSII (ref. 7), whereas PsbW is present only in higher plants and algae but not in cyanobacteria²⁰. In red algal PSII (CcPSII from *Cyanidium caldarium*)²¹, a subunit found at a location adjacent to PsbI and named ‘chain S’ may correspond to spinach PsbW. PsbZ is the only subunit with two transmembrane helices and its N and C termini are both positioned on the luminal surface, whereas the other 11 subunits all have a single transmembrane helix. The N termini of the PsbE, PsbF, PsbL, PsbJ and PsbH subunits are located at the stromal surface, whereas PsbI, PsbK, PsbM, PsbTc, PsbW and PsbX assume a reverse topology with their N termini positioned on the luminal side. These small intrinsic subunits are involved in dimerization of the core complex (PsbTc, PsbL and PsbM), stabilization of the core (PsbK, PsbJ, PsbE, PsbF and PsbX), mediating the association of peripheral antenna complexes with the core complex (PsbW, PsbZ and PsbH), and binding cytochrome b_{559} to protect PSII from photodamage (PsbE and PsbF)²².

At the outer region of the core, one LHCII trimer and one CP26 monomer flank the side near CP43, and one CP29 monomer is associated with CP47 on the other side (Fig. 2a, c). Within each monomer of the dimeric supercomplex, a large number of cofactors have been located, as summarized in Extended Data Table 1. They include 105 chlorophyll (Chl) molecules, 28 β -carotene and xanthophylls, one haem, one Mn_4CaO_5 cluster, one plastoquinone and numerous lipid molecules. Whereas the core subunits (CP43, CP47, D1 and D2) bind

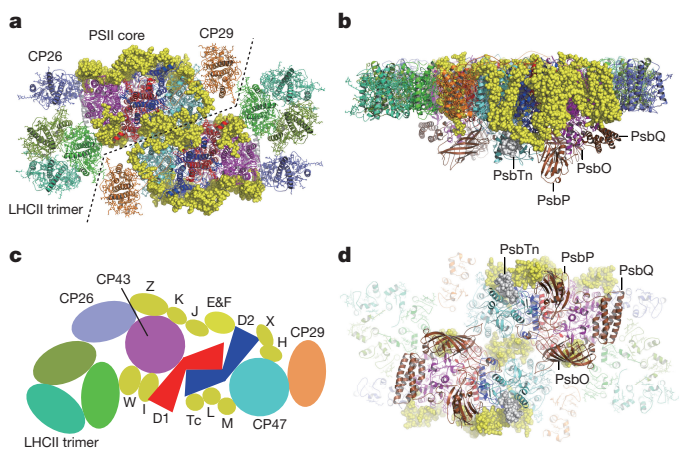


Figure 2 | Overall architecture of spinach PSII-LHCII supercomplex.

a, **b**, Structure of the spinach C_2S_2 -type PSII-LHCII supercomplex. **a**, View from the stromal side along membrane normal. **b**, Side view along membrane plane. Dashed lines indicate estimated interfacial regions between the two monomers. The major components are shown as cartoon and stick models in different colours and the 12 small intrinsic subunits are shown as yellow sphere models. **c**, Cartoon diagram showing the assignment of membrane-embedded subunits of the supercomplex. Only one monomer is shown and the colour codes are consistent with those in **a**, **d**. **d**, Lumen-exposed regions of the supercomplex. The view is along the membrane normal from the luminal side.

only Chl *a* and β -carotene molecules, LHCII, CP29 and CP26 contain both Chl *a* and Chl *b*, and three different xanthophylls (lutein, neoxanthin and violaxanthin)^{4,12,23}.

The extrinsic subunits

On the luminal side of the core complex, we have located the binding sites for four extrinsic subunits, namely PsbO, PsbP, PsbQ and PsbTn (Fig. 3a, b and Extended Data Fig. 4a). Among them, PsbO, PsbP and PsbQ form a triangular crown-like structure encircling the luminal domain of CP43 and the C-terminal tail of D1 (Fig. 3b, c). These parts of CP43 and D1 are directly involved in coordinating and providing shields for the Mn_4CaO_5 cluster, the oxygen-evolving centre responsible for splitting water into oxygen and protons⁷. By interacting with the two Mn_4CaO_5 -binding regions, the heterotrimeric PsbO–PsbP–PsbQ complex serves to optimize the efficiency of oxygen evolution in PSII under physiological conditions²⁴. Spinach PsbO has a characteristic β -barrel structure similar to its homologue in TvPSII with an r.m.s.d. of α -carbon atoms at 1.42 Å. Near the $\beta 7$ – $\beta 8$ loop region of PsbO, PsbP binds in a canyon between the luminal domains of CP43 and CP47 (Fig. 3b). A short loop between Asp137 and Glu140 of PsbP stabilizes the C-terminal tails of D1 and D2 (Fig. 3c). The binding site of PsbP partly overlaps with those of PsbV and PsbU in TvPSII (ref. 7) and CcPSII (ref. 21) (Extended Data Fig. 4b). On the other side, the N-terminal region of spinach PsbO binds to the four-helix bundle domain of PsbQ, and PsbQ simultaneously interacts with the luminal domain of CP43 (Fig. 3a). In CcPSII (ref. 21), a PsbQ' protein resembling spinach PsbQ was found in a similar location (Extended Data Fig. 4b). Curiously, the elongated N-terminal region of spinach PsbQ reaches out approximately 50 Å away from the four-helix bundle and binds to a long loop of PsbP between Lys90 and Ala111. The interaction between PsbP and PsbQ was previously detected through a crosslinking method combined with mass spectrometry^{25,26}. Marked conformational changes occur in the flexible regions of PsbP and PsbQ when they bind to PSII core subunits (Extended Data Fig. 4c, d). Besides them, the smallest subunit of plant PSII with unknown function, PsbTn (nuclear encoded PsbT subunit)²⁷, intercalates between the luminal domain of CP47 and the C-terminal region of PsbE, serving as a bridge between them (Fig. 3b).

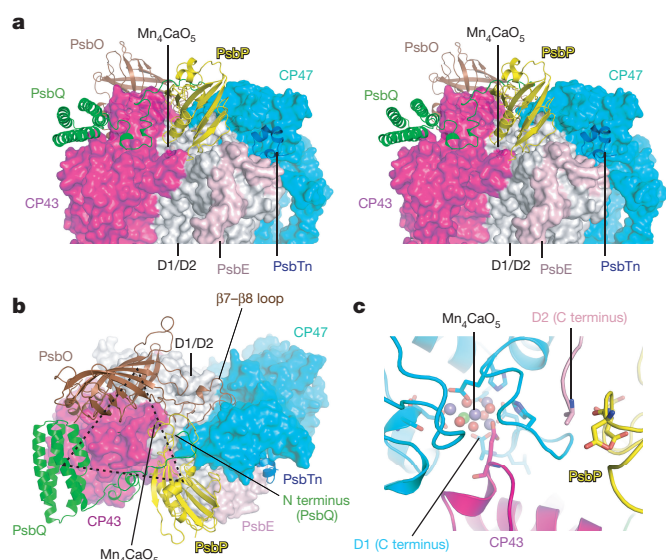


Figure 3 | The extrinsic subunits of spinach PSII. **a**, Cartoon representation of the extrinsic subunits surrounding the oxygen-evolving centre. The figure is shown as a stereo pair viewed along the membrane plane. The peripheral antenna complexes are omitted for clarity. The four extrinsic subunits (PsbO, PsbP, PsbQ and PsbTn) are shown as ribbon models; the intrinsic subunits of the PSII core are presented as surface models. **b**, The location of the four extrinsic subunits viewed from the luminal side. The dashed-line triangle indicates the PsbO–PsbP–PsbQ tertiary complex, which forms a crown-like structure encircling the luminal domain of CP43 and D1. **c**, The role of a loop from PsbP in stabilizing the C-terminal regions of D1 and D2. The protein backbones are shown as ribbons; the Mn_4CaO_5 cluster is shown as a sphere model; and the amino acid residues involved in coordinating the cluster are presented as stick models. The segment of loop between Asp137 and Glu140 in PsbP involved in contacting the C-terminal tails of D1 and D2 subunits is highlighted as a stick model.

Structures of peripheral antennae

The cryo-EM structure of LHCII within the supercomplex is nearly identical to the previous crystal structure of isolated LHCII¹² (Extended Data Fig. 5). The densities and structures of CP29 and CP26 are shown in Extended Data Figs 6 and 7. The binding sites and identities of chlorophylls and carotenoids in LHCII, CP29 and CP26 are summarized in Extended Data Table 2. The chromophore identities were assigned according to the appearance of cryo-EM densities (Extended Data Fig. 8a), previous high-resolution crystal structures of spinach LHCII (ref. 12) and CP29 (ref. 14), and the functional architecture of CP26 (ref. 23).

For CP29, three carotenoid and thirteen chlorophyll binding sites were located. Despite their overall similarity, there are evident differences between the cryo-EM structure of full-length CP29 and the crystal structure of CP29 without its N-terminal domain¹⁴ (Extended Data Fig. 6a, b). The long N-terminal region (87 amino acid residues) of CP29 was unobserved in the previous crystal structure owing to its high flexibility and proteolysis during crystallization¹⁴. An earlier work using electron paramagnetic resonance approaches suggested that this region is potentially structured²⁸. The cryo-EM structure of CP29 shows that this region forms two motifs with irregular coil structures (motifs I and II) (Extended Data Fig. 6b). Motif I (Pro12–Lys41) superposes well with the corresponding N-terminal region of LHCII (Extended Data Fig. 6c). A chlorophyll density resembling Chl *b*601 of LHCII is observed in this region (Extended Data Fig. 6c, d). It is located near Chl *a*611 with a Mg-to-Mg distance of 12.0 Å and is tentatively assigned as a Chl *a*. Its central ligand is contributed by the carbonyl of Trp14 from the N-terminal region of CP29. Motif II (Pro42–Phe87) forms an L-shaped structure containing an approximately 40 Å-long hairpin loop (Pro42–Ser72) running nearly parallel to the stromal surface,

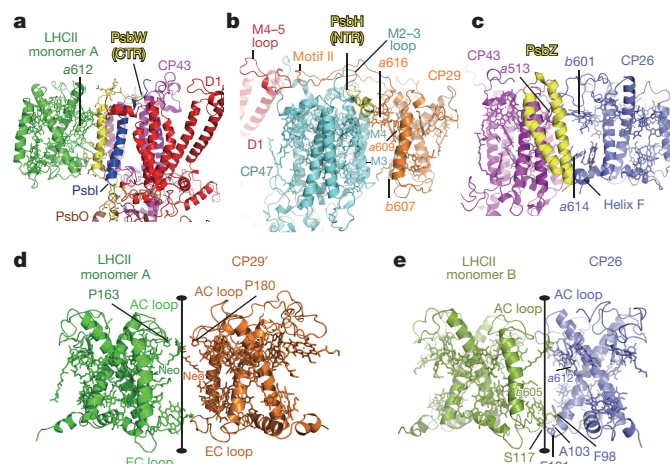


Figure 4 | Antenna-core and antenna-antenna interactions in the PSII-LHCII supercomplex. **a**, The interface between LHCII and CP43/D1. Their interactions are mediated by PsbW and PsbI. PsbW (CTR), C-terminal region of the PsbW subunit. **b**, Interactions between CP29 and CP47/D1. M4–5 loop, loop region between the fourth and fifth transmembrane helices of D1; M3 and M4, third and fourth transmembrane helices of CP47, respectively; M2–3 loop, loop between the second and third transmembrane helices of CP47; PsbH (NTR), N-terminal region of the PsbH subunit. **c**, Interactions between CP26 and CP43/PsbZ. **d**, Contacts between LHCII and CP29'. **e**, CP26–LHCII interactions.

and a short hairpin loop (Ala73–Phe87) beneath the long hairpin (Extended Data Fig. 6b). Strikingly, a new chlorophyll-binding site is found attached to the short hairpin in motif II and located at the interfacial region facing CP47 (Extended Data Fig. 6e). This chlorophyll is coordinated by the backbone carbonyl of Leu80 and is tentatively assigned as a Chl *a* with its binding site named 616. Chl *a*616 in CP29 superposes partly with Chl *a*617 in Lhca3 and Lhca4 of the PSI–LHCI supercomplex^{29,30} (Extended Data Fig. 6f), suggesting that these chlorophylls may serve similar roles as interfacial chlorophylls facilitating energy transfer between two adjacent antenna complexes.

For CP26 in the PSII-LHCII supercomplex, thirteen chlorophyll binding sites and three carotenoid binding sites were observed (Extended Data Table 2 and Extended Data Fig. 7). Among them, Chls *b*601, *a*604, *b*607 and *b*608 were not predicted by the previous functional study²³, as their central ligands are contributed by either a protein backbone carbonyl group (601) or potential water molecules (604, 607 and 608). Their identities are tentatively assigned according to their similarity to the corresponding sites in LHCII. The carotenoids are assigned as two luteins (L1 and L2) and one neoxanthin (N1). Although the L1 site is mainly occupied by lutein, the L2 site can also accept violaxanthin as well as lutein²³, and the N1 site binds neoxanthin³¹.

Antenna-PSII core assembly

Efficient transfer of excitation energy from the peripheral antenna complexes to the PSII core relies on their specific non-covalent interactions. The association of the LHCII trimer with the core complex is mediated by PsbW (Fig. 4a) and the interfacial lipid molecules (Extended Data Fig. 8b). On the stromal side, the LHCII trimer binds to PsbW through hydrophobic interactions between its Chl *a*611–*a*612 pair and Trp117/Phe121 from PsbW. On the luminal side, Asn88 from LHCII is hydrogen-bonded to Trp107 and Asn103 on PsbW. Meanwhile, Leu84_{LHCII} forms van der Waals contacts with Trp107_{PsbW}. To further connect LHCII with the core, the transmembrane helix of PsbW forms extensive hydrophobic interactions with PsbI located on the side opposite to the LHCII binding site. PsbI simultaneously interacts with the first transmembrane helix of the D1 subunit and PsbW. In addition, the N-terminal region of PsbW extends to the luminal surface and interacts with PsbO and D1, while its C-terminal region is located on the stromal

surface and binds to the loop region between the fourth and fifth transmembrane helices of CP43. In *PsbW* knockout plants, the PSII–LHCII supercomplexes are destabilized and could not be detected²⁰. The location of *PsbW* at the interface between the LHCII trimer and the core complex explains its essential role for the formation of the PSII–LHCII supercomplex. In addition, numerous lipid-like densities are found at the interfaces between *PsbW* and LHCII and between CP43 and LHCII, enhancing the association of LHCII with the core complex (Extended Data Fig. 8b).

Both CP29 and CP26 contribute to the proper assembly and stability of the PSII–LHCII supercomplex^{32,33}. The transmembrane domain of CP29 directly associates with that of CP47 through hydrophobic interactions. Notably, Chl *a*616, the Chl *a*603 and *a*609 pair and Chl *b*607 from CP29 make several contacts with hydrophobic residues from the third and fourth transmembrane helices of CP47 (Fig. 4b). Above Chl *a*616, the long hairpin loop in motif II of CP29 binds to the stromal surface of CP47 through hydrogen bonds and van der Waals interactions. This loop further reaches out to contact the Thr228–Asn230 region on a surface loop of the D1 subunit. Moreover, the elongated N-terminal region of the *PsbH* subunit forms a loop–helix–loop structure and intercalates between motif II of CP29 and the loop between the second and third transmembrane helices of CP47. Thus, *PsbH* secures the interactions between CP29 and CP47 (Fig. 4b).

CP26 interacts specifically with CP43 and *PsbZ* through its N-terminal and C-terminal regions as well as the chlorophylls (Chl *b*601 and Chl *a*614) bound to these two regions (Fig. 4c). The N-terminal region of CP26 (Pro36–Leu39) and Chl *b*601 (coordinated by Phe34 in this region) form van der Waals contacts and hydrophobic interactions with Chl *a*513 from CP43. The C-terminal region of CP26 contains a short amphipathic α -helix between Leu230 and Ile234 (named helix F), and a loop named the DF loop (Pro225–Asn229) preceding helix F. The DF loop forms van der Waals contacts with Phe182 from CP43, and is further bridged to the luminal surface of CP43 through interfacial lipid molecules (Extended Data Fig. 8b). Helix F is sandwiched between helix A_{CP26} and the second α -helix of *PsbZ* (Fig. 4c). It binds to the C-terminal region of *PsbZ* through hydrophobic interactions and a hydrogen bond (Leu231_{CP26}–Ser59_{PsbZ}). The second helix of *PsbZ* is positioned at the interface between CP26 and CP43, and the first helix binds to CP43 on its membrane-facing surface and interacts closely with the second helix so as to provide rigid support for it. Thereby, *PsbZ* reinforces the association of CP26 with CP43. When *psbZ* is knocked out, the content of CP26 protein decreases markedly and the formation of the PSII–LHCII supercomplex is deficient^{34–36}.

Interactions between LHCII, CP29 and CP26

The LHCII trimer serves as a bridge connecting CP26 with CP29 from the adjacent monomer (CP29') of the dimeric supercomplex (Fig. 2a). The three monomers of the LHCII trimer within the supercomplex are not equivalent, as they are surrounded by different neighbours. Two of them, monomers A and B, interact with CP29' and CP26, respectively, whereas the third monomer (C) is located at the peripheral region (contacting the moderately associated LHCII in the larger C₂S₂M₂ supercomplex¹⁶). As shown in Fig. 4d, e, monomers A and B are related to the adjacent CP29' and CP26, respectively, through pseudo-C₂ symmetry running through their interfaces. Monomer A of LHCII forms several contacts with CP29' on both the stromal and luminal sides (Fig. 4d). On the stromal surface, Pro163 from the AC loop (between helices A and C) region of LHCII interacts with the trimethylcyclohexane-1,3-diol head group of neoxanthin from CP29'. Meanwhile, the neoxanthin from LHCII contacts Pro180 from the AC loop region of CP29'. On the luminal side, the EC loop and Chl *b*605 (bound to Val119 in this region) of LHCII associate with the Gly137–Pro141 segment of the EC loop of CP29'. The interactions between LHCII and CP29' are further strengthened by the lipid-like molecules found in the interfacial gaps (Extended Data Fig. 8b).

For monomer B of the LHCII trimer, the region between Ser160 and Pro163 in the AC loop binds to the Pro172–Gly174 region in the AC loop of CP26 (Fig. 4e). Neoxanthin from LHCII monomer B is in contact with Chl *a*612 of CP26. On the luminal side, Ser117 from the EC loop of LHCII forms a hydrogen bond with the backbone amide of Ala103 from the BE loop of CP26. Moreover, Chl *b*605 makes van der Waals contacts with Phe98 and Phe101 from Helix B of CP26. Evidently, the AC loop, neoxanthin molecule and EC loop (and the associated Chl *b*605) are important components for LHCII to bind and recognize CP26 and CP29' within the supercomplex.

Insights into energy transfer pathways

Through the interfacial chlorophyll pairs located between the peripheral and core antenna complexes, excitation energy can be transferred from LHCII, CP26 and CP29 to CP43 or CP47. The LHCII trimer contains three Chl *b*-rich clusters located at its monomer–monomer interface¹², and two of these clusters are connected to CP26 and CP29' (Fig. 5a). As Chl *b* has a higher energy level than Chl *a*, the energy transfer between LHCII and CP26 (or between LHCII and CP29') may flow from the Chl *b*-rich regions of LHCII to adjacent Chl *a*-rich regions in CP26 (or CP29'). For energy transfer between LHCII and CP29', Chl *b*605_{monomerA/LHCII} and Chl *a*604/*b*606_{CP29'} from the luminal layer form the closest inter-complex pair, with a Mg-to-Mg distance (D_{centre}) of 17.7/18.6 Å, while Chl *b*608_{monomerA/LHCII} and Chl *b*608_{CP29'} at the stromal layer are connected with D_{centre} at 23.0 Å (Fig. 5b). Thus, the Chl *b*-rich region of CP29' joins with that of LHCII, facilitating energy transfer between two adjacent monomers of the dimeric supercomplex, presumably through Chl *b*605_{monomerA/LHCII} to Chl *a*604/*b*606_{CP29'} and *b*608_{monomerA/LHCII} to *b*608_{CP29'} pathways. For energy transfer between LHCII and CP26, Chl *b*608 from monomer B of the LHCII trimer is connected to Chl *a*612 and Chl *a*610 from CP26 with D_{centre} at 21.8 and 21.9 Å, respectively, while Chl *b*605 from the luminal layer of LHCII may transfer its excitation energy to Chl *a*604 of CP26 at 19.3 Å D_{centre} (Fig. 5c).

The lowest energy-state chlorophylls in LHCII were attributed to the Chl *a*610, *a*612 and *a*611 cluster and are known as the terminal emitter domain^{37–39}. The excitation energy equilibrated within the LHCII trimer will be focused on this cluster⁴⁰. In the supercomplex, the terminal emitter from monomer A of the LHCII trimer may transfer its energy to Chl *a*506_{CP43}, which is located in a favourable orientation (nearly parallel) and distance (17.1 Å D_{centre}) with respect to Chl *a*611_{monomerA/LHCII} (Fig. 5d). Below Chl *a*611, Chl *a*614 from the luminal layer of LHCII is connected to Chl *a*501_{CP43} at 25.1 Å D_{centre} . These two pathways form the bases of energy transfer between LHCII and the core complex. In the absence of minor antenna complexes, LHCII can transfer energy directly to the core complexes⁴¹, but the functional connection between LHCII and the PSII core is severely impaired in minor-antenna knockout mutant plants⁴².

For energy transfer between CP29 and CP47, Chl *a*616_{CP29} is sandwiched between Chl *a*609_{CP29} and Chl *a*616_{CP47} at 9.3 and 14.4 Å D_{centre} , respectively (Fig. 5e). The closest edge-to-edge distance (D_{edge}) between Chl *a*616_{CP29} and Chl *a*609_{CP29}/*a*616_{CP47} is 3.4/4.2 Å, indicating that these chlorophylls form strongly coupled pairs. The interstitial position of Chl *a*616_{CP29} makes it a crucial linker, relaying the transfer of excitation energy from CP29 to CP47. In addition, Chl *a*603 from CP29 is directly connected to Chl *a*610_{CP47} at 18.6 Å D_{centre} . At the luminal layer, energy may be transferred from Chl *b*607_{CP29} to Chl *a*607_{CP47} at 19.1 Å D_{centre} . Alternatively, Chl *b*607_{CP29} may transfer its energy to Chl *a*603–*a*609_{CP29}, and the energy may be further relayed by Chl *a*616_{CP29} to Chl *a*616_{CP47}. Among these potential pathways, Chl *a*616_{CP29} to Chl *a*616_{CP47} is likely to be the most efficient energy transfer pathway between CP29 and CP47, as these two chlorophylls are the most closely paired at the interface.

CP26 interacts closely with CP43 and energy transfer between them may occur through multiple potential pathways (Fig. 5f). Chl *a*611_{CP26} forms a strongly coupled pair with the red-most Chl *a*612 (ref. 23) and

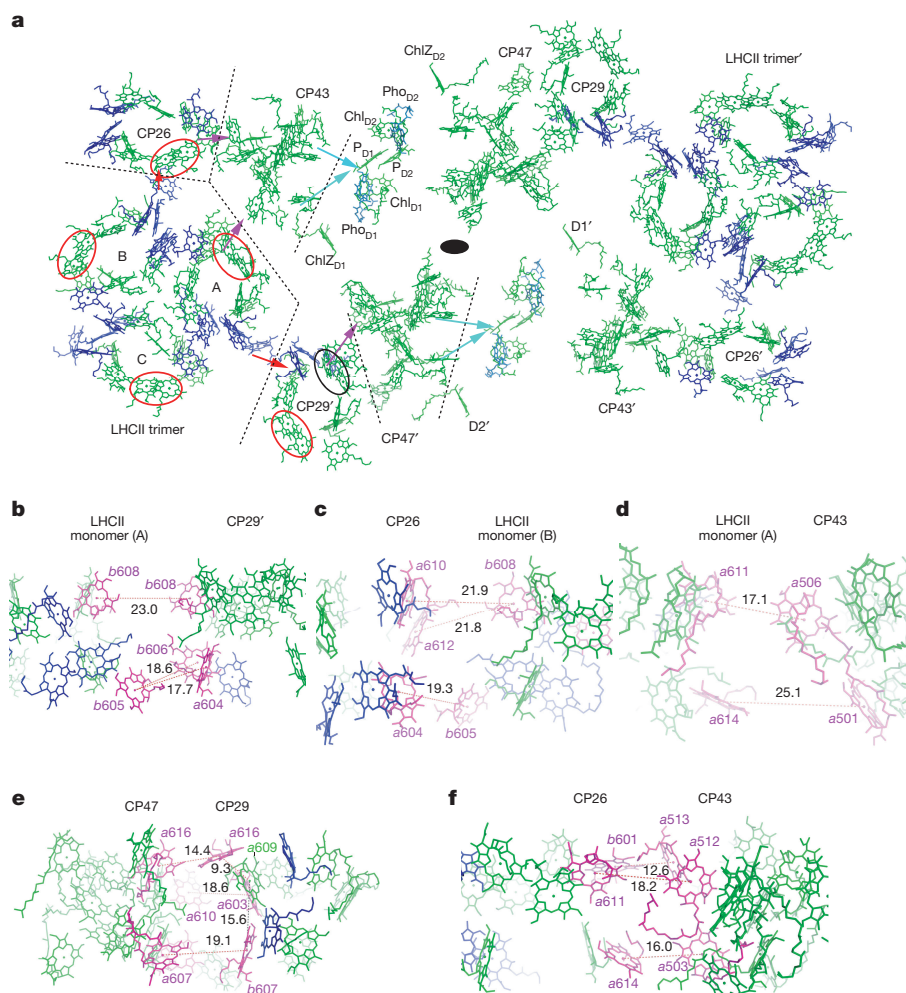


Figure 5 | Energy transfer pathways from antenna complexes to reaction centres. **a**, Distribution pattern of chlorophylls within the PSII-LHCII supercomplex. Chl *a* and Chl *b* are coloured green and blue, respectively. Arrows indicate potential energy transfer pathways between LHCII and CP29'/CP26 (red), LHCII/CP29'/CP26 and core antennae (CP43/CP47') (magenta), and CP43/CP47' and the reaction centre in D1/D2' (cyan). The five red ovals indicate the potential energy-quenching sites around the Chl *a*611–Chl *a*612 pairs, and the black oval indicates the other potential quenching site around the Chl *a*603–Chl *a*609 pair in CP29'. These

quenching sites might be activated under high-light conditions so as to dissipate harmful excess energy. The dashed lines indicate the approximate boundaries of each individual complex. **b–d**, The interfacial chlorophylls supporting energy transfer between LHCII and CP29' (**b**), LHCII and CP26 (**c**), and LHCII and CP43 (**d**). **e, f**, The chlorophylls at the interfaces between CP29 and CP47 (**e**), and CP26 and CP43 (**f**). The numbers near the dashed lines indicate the Mg-to-Mg distances (Å) between two adjacent chlorophylls. The interfacial chlorophylls are highlighted in magenta.

this pair is probably the terminal emitter in CP26. Chl *a*611 is connected to Chl *a*512_{CP43} at 18.2 Å D_{centre} and to Chl *a*513_{CP43} at 17.2 Å D_{centre} . Meanwhile, Chl *b*601_{CP26} is coupled to Chl *a*513_{CP43} at 4.8 Å D_{edge} (with D_{centre} at 12.6 Å) and is also connected to Chl *a*512_{CP43} at 19.2 Å D_{centre} . At the luminal layer, the excitation energy from Chl *a*614_{CP26} may be absorbed primarily by Chl *a*503_{CP43} at 16.0 Å D_{centre} . Thus, the energy transmitted from CP26 will be received by the Chl *a*513–Chl *a*512 pair at the stromal layer, or by Chl *a*503 at the luminal layer of CP43.

When the excitation energy from the peripheral antenna complexes has been collected by the core antenna complexes, subsequent energy transfer from CP47 or CP43 to the P680 special pair occurs through the Chl *a* network located within CP43, CP47, D1 and D2, as indicated in Fig. 5a. Under high-light conditions, clusters of pigment molecules within the major and minor LHCII may serve as non-photochemical quenching sites that dissipate harmful excess energy as heat^{43–45}. The potential quenching sites within the supercomplex are mainly located at or near the interfaces between adjacent antenna complexes (Fig. 5a). These locations are ideal for them to intercept and dissipate excess energy before it reaches the reaction centre. Recently, biophysical modelling studies have yielded preliminary information about the

kinetics of light harvesting in PSII-LHCII supercomplexes^{46,47}. Now, the cryo-EM structure of the spinach PSII-LHCII supercomplex provides a detailed framework of its highly sophisticated pigment network and enables a deeper understanding of the kinetics and regulation of light-harvesting processes within the supercomplex.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 11 March; accepted 19 April 2016.

Published online 18 May; corrected online 1 June 2016

(see full-text HTML version for details).

1. Nelson, N. & Junge, W. Structure and energy transfer in photosystems of oxygenic photosynthesis. *Annu. Rev. Biochem.* **84**, 659–683 (2015).
2. Shen, J.-R. The structure of photosystem II and the mechanism of water oxidation in photosynthesis. *Annu. Rev. Plant Biol.* **66**, 23–48 (2015).
3. Vinyard, D. J., Ananyev, G. M. & Dismukes, G. C. Photosystem II: The reaction center of oxygenic photosynthesis. *Annu. Rev. Biochem.* **82**, 577–606 (2013).
4. Pan, X., Liu, Z., Li, M. & Chang, W. Architecture and function of plant light-harvesting complexes II. *Curr. Opin. Struct. Biol.* **23**, 515–525 (2013).
5. Barros, T. & Kühlbrandt, W. Crystallisation, structure and function of plant light-harvesting Complex II. *Biochim. Biophys. Acta* **1787**, 753–772 (2009).

6. Croce, R. & van Amerongen, H. Natural strategies for photosynthetic light harvesting. *Nature Chem. Biol.* **10**, 492–501 (2014).
7. Umena, Y., Kawakami, K., Shen, J.-R. & Kamiya, N. Crystal structure of oxygen-evolving photosystem II at a resolution of 1.9 Å. *Nature* **473**, 55–60 (2011).
8. Suga, M. *et al.* Native structure of photosystem II at 1.95 Å resolution viewed by femtosecond X-ray pulses. *Nature* **517**, 99–103 (2015).
9. Rhee, K.-H., Morris, E. P., Barber, J. & Kühlbrandt, W. Three-dimensional structure of the plant photosystem II reaction centre at 8 Å resolution. *Nature* **396**, 283–286 (1998).
10. Hankamer, B., Morris, E. P. & Barber, J. Revealing the structure of the oxygen-evolving core dimer of photosystem II by cryoelectron crystallography. *Nature Struct. Mol. Biol.* **6**, 560–564 (1999).
11. Hankamer, B., Morris, E., Nield, J., Gerle, C. & Barber, J. Three-dimensional structure of the photosystem II core dimer of higher plants determined by electron microscopy. *J. Struct. Biol.* **135**, 262–269 (2001).
12. Liu, Z. *et al.* Crystal structure of spinach major light-harvesting complex at 2.72 Å resolution. *Nature* **428**, 287–292 (2004).
13. Standfuss, J., Terwisscha van Scheltinga, A. C., Lamborghini, M. & Kühlbrandt, W. Mechanisms of photoprotection and nonphotochemical quenching in pea light-harvesting complex at 2.5 Å resolution. *EMBO J.* **24**, 919–928 (2005).
14. Pan, X. *et al.* Structural insights into energy regulation of light-harvesting complex CP29 from spinach. *Nature Struct. Mol. Biol.* **18**, 309–315 (2011).
15. Nield, J. & Barber, J. Refinement of the structural model for the Photosystem II supercomplex of higher plants. *Biochim. Biophys. Acta* **1757**, 353–361 (2006).
16. Caffarri, S., Kouril, R., Kereiche, S., Boekema, E. J. & Croce, R. Functional architecture of higher plant photosystem II supercomplexes. *EMBO J.* **28**, 3052–3063 (2009).
17. Drop, B. *et al.* Light-harvesting complex II (LHCII) and its supramolecular organization in *Chlamydomonas reinhardtii*. *Biochim. Biophys. Acta* **1837**, 63–72 (2014).
18. Guskov, A. *et al.* Cyanobacterial photosystem II at 2.9-Å resolution and the role of quinones, lipids, channels and chloride. *Nature Struct. Mol. Biol.* **16**, 334–342 (2009).
19. Ferreira, K. N., Iverson, T. M., Maghlaoui, K., Barber, J. & Iwata, S. Architecture of the photosynthetic oxygen-evolving center. *Science* **303**, 1831–1838 (2004).
20. Garcia-Cerdán, J. G. *et al.* The PsbW protein stabilizes the supramolecular organization of photosystem II in higher plants. *Plant J.* **65**, 368–381 (2011).
21. Ago, H. *et al.* Novel features of eukaryotic photosystem II revealed by its crystal structure analysis from a red alga. *J. Biol. Chem.* **291**, 5676–5687 (2016).
22. Shi, L. X. & Schroder, W. P. The low molecular mass subunits of the photosynthetic supercomplex, photosystem II. *Biochim. Biophys. Acta* **1608**, 75–96 (2004).
23. Ballottari, M., Mozzo, M., Croce, R., Morosinotto, T. & Bassi, R. Occupancy and functional architecture of the pigment binding sites of photosystem II antenna complex Lhcb5. *J. Biol. Chem.* **284**, 8103–8113 (2009).
24. Bricker, T. M., Roose, J. L., Fagerlund, R. D., Frankel, L. K. & Eaton-Rye, J. J. The extrinsic proteins of Photosystem II. *Biochim. Biophys. Acta* **1817**, 121–142 (2012).
25. Ido, K. *et al.* Cross-linking evidence for multiple interactions of the PsbP and PsbQ proteins in a higher plant photosystem II supercomplex. *J. Biol. Chem.* **289**, 20150–20157 (2014).
26. Mummadiiseti, M. P. *et al.* Use of protein cross-linking and radiolytic footprinting to elucidate PsbP and PsbQ interactions within higher plant Photosystem II. *Proc. Natl Acad. Sci. USA* **111**, 16178–16183 (2014).
27. Kapazoglou, A., Sagliocco, F. & Dure, L. PSII-T, a new nuclear encoded luminal protein from photosystem II: targeting and processing in isolated chloroplasts. *J. Biol. Chem.* **270**, 12197–12202 (1995).
28. Shabestari, M. H., Wolfs, C. J. A. M. & Spruijt, R. B. van Amerongen, H. & Huber, M. Exploring the structure of the 100 amino-acid residue long N-terminus of the plant antenna protein CP29. *Biophys. J.* **106**, 1349–1358 (2014).
29. Qin, X., Suga, M., Kuang, T. & Shen, J. R. Photosynthesis. Structural basis for energy transfer pathways in the plant PSI-LHCII supercomplex. *Science* **348**, 989–995 (2015).
30. Mazar, Y., Borovikova, A. & Nelson, N. The structure of plant photosystem I super-complex at 2.8 Å resolution. *eLife* **4**, e07433 (2015).
31. Caffarri, S., Passarini, F., Bassi, R. & Croce, R. A specific binding site for neoxanthin in the monomeric antenna proteins CP26 and CP29 of Photosystem II. *FEBS Lett.* **581**, 4704–4710 (2007).
32. Yakushevska, A. E. *et al.* The structure of photosystem II in *Arabidopsis*: localization of the CP26 and CP29 antenna complexes. *Biochemistry* **42**, 608–613 (2003).
33. de Bianchi, S. *et al.* *Arabidopsis* mutants deleted in the light-harvesting protein Lhcb4 have a disrupted photosystem II macrostructure and are defective in photoprotection. *Plant Cell* **23**, 2659–2679 (2011).
34. Ruf, S., Biehler, K. & Bock, R. A small chloroplast-encoded protein as a novel architectural component of the light-harvesting antenna. *J. Cell Biol.* **149**, 369–378 (2000).
35. Baena-González, E., Gray, J. C., Tyystjärvi, E., Aro, E.-M. & Mäenpää, P. Abnormal regulation of photosynthetic electron transport in a chloroplast ycf9 inactivation mutant. *J. Biol. Chem.* **276**, 20795–20802 (2001).
36. Swiatek, M. *et al.* The chloroplast gene ycf9 encodes a photosystem II (PSII) core subunit, PsbZ, that participates in PSII supramolecular architecture. *Plant Cell* **13**, 1347–1368 (2001).
37. Remelli, R., Varotto, C., Sandona, D., Croce, R. & Bassi, R. Chlorophyll binding to monomeric light-harvesting complex. A mutation analysis of chromophore-binding residues. *J. Biol. Chem.* **274**, 33510–33521 (1999).
38. Rogl, H. & Kühlbrandt, W. Mutant trimers of light-harvesting complex II exhibit altered pigment content and spectroscopic features. *Biochemistry* **38**, 16214–16222 (1999).
39. Novoderezhkin, V. I., Palacios, M. A., van Amerongen, H. & van Grondelle, R. Excitation dynamics in the LHCII complex of higher plants: modeling based on the 2.72 Å crystal structure. *J. Phys. Chem. B* **109**, 10493–10504 (2005).
40. Novoderezhkin, V., Marin, A. & van Grondelle, R. Intra- and inter-monomeric transfers in the light harvesting LHCII complex: the Redfield-Förster picture. *Phys. Chem. Chem. Phys.* **13**, 17093–17103 (2011).
41. Sun, R. *et al.* Direct energy transfer from the major antenna to the photosystem II core complexes in the absence of minor antennae in liposomes. *Biochim. Biophys. Acta* **1847**, 248–261 (2015).
42. Dall'Osto, L., Ünlü, C., Cazzaniga, S. & van Amerongen, H. Disturbed excitation energy transfer in *Arabidopsis thaliana* mutants lacking minor antenna complexes of photosystem II. *Biochim. Biophys. Acta* **1837**, 1981–1988 (2014).
43. Mozzo, M., Passarini, F., Bassi, R., van Amerongen, H. & Croce, R. Photoprotection in higher plants: The putative quenching site is conserved in all outer light-harvesting complexes of Photosystem II. *Biochim. Biophys. Acta* **1777**, 1263–1267 (2008).
44. Ahn, T. K. *et al.* Architecture of a charge-transfer state regulating light harvesting in a plant antenna protein. *Science* **320**, 794–797 (2008).
45. Pascal, A. A. *et al.* Molecular basis of photoprotection and control of photosynthetic light-harvesting. *Nature* **436**, 134–137 (2005).
46. Caffarri, S., Broess, K., Croce, R. & van Amerongen, H. Excitation energy transfer and trapping in higher plant Photosystem II complexes with different antenna sizes. *Biophys. J.* **100**, 2094–2103 (2011).
47. Bennett, D. I. G., Amarnath, K. & Fleming, G. R. A. Structure-based model of energy transfer reveals the principles of light harvesting in photosystem II supercomplexes. *J. Am. Chem. Soc.* **135**, 9164–9173 (2013).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank J. P. Zhang and X. L. Zhao for their assistance in preparing thylakoid samples. Cryo-EM data collection was carried out at the Center for Biological Imaging, Core Facilities for Protein Science at the Institute of Biophysics (IBP), Chinese Academy of Sciences (CAS), and at the National Center for Protein Science Shanghai (NCPSS), Institute of Biochemistry and Cell Biology, Shanghai Institutes for Biological Sciences/Shanghai Science Research Center, Chinese Academy of Sciences, Shanghai, China. We thank X. J. Huang, G. Ji, W. Ding, F. Sun, and other staff members at the Center for Biological Imaging (IBP, CAS); L. L. Kong, X. Y. Shi, Y. N. He, J. P. Ding, and M. Lei for their support during data collection; X. B. Liang and X. M. An for support in organizing data collection trips; F. L. Zhang, J. Zhou, and Y. Li for support in measuring the oxygen evolution activity; L. L. Niu and X. Ding for mass spectrometry; J. H. Li for assistance in fluorescence measurement; R. Bassi, A. Pinnola and R. Croce for sharing experiences in purifying plant PSII-LHCII supercomplexes; and Y. Xiang for advice on cryo-EM sample preparation and structure refinement. The project was funded by National 973 project grant 2011CBA00900, the Strategic Priority Research Program of CAS (XDB08020302) and National Natural Science Foundation of China (31570724, 31270793 and 31170703). Z.L. and X.Z. received scholarships from the 'National Thousand (Young) Talents Program' from the Office of Global Experts Recruitment in China.

Author Contributions X.W., X.S., P.C. and X.L. purified the spinach PSII-LHCII supercomplex; M.L. and P.C. characterized the spectroscopic features, protein and pigment contents, and oxygen-evolving activity of the samples; X.W., X.S. and X.Z. collected and processed cryo-EM data; X.Z. reconstructed the 3.2 Å resolution map and supervised cryo-EM structure determination; X.W. and Z.L. built and refined the structure model; X.W., M.L., X.Z. and Z.L. analysed the structure; Z.L. and W.C. conceived and coordinated the project; and the manuscript was written by X.W., M.L., X.Z. and Z.L.

Author Information The cryo-EM map of the spinach PSII-LHCII supercomplex has been deposited in the Electron Microscopy Data Bank with accession code EMD-6617. The corresponding structure model has been deposited in the Protein Data Bank under accession code 3JCU. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to X.Z. (xzzhang@ibp.ac.cn, for cryo-EM data collection, processing and structure determination), M.L. (meli@moon.ibp.ac.cn, for sample preparation and characterization) or Z.L. (liuzf@sun5.ibp.ac.cn, for structure and function of the PSII-LHCII supercomplex).

Reviewer Information Nature thanks Roberta Croce, Jian-Ren Shen and Thomas Walz for their contribution to the peer review of this work.

METHODS

No statistical methods were used to predetermine sample size, the experiments were not randomized and the investigators were not blinded to allocation during experiments and outcome assessment.

Purification and characterization of spinach PSII–LHCII supercomplex. Grana membranes were prepared from spinach leaves as described previously¹⁶. For purification of PSII–LHCII supercomplexes, 500 µg (in chlorophyll) of spinach grana membrane was washed once with 1 mM EDTA, 10 mM HEPES (pH 7.5) and then centrifuged at 16,000g for 10 min. Subsequently, the pellets were suspended in 10 mM HEPES (pH 7.5) and then solubilized at 0.5 mg ml^{−1} Chl by adding an equal volume of 0.6% dodecyl- α -D-maltoside (α -DDM) in 10 mM HEPES (pH 7.5) and vortexing for 1 min. The solubilized sample was centrifuged at 16,000g for 10 min to remove pellets and the supernatant was fractionated through sucrose-density gradient ultracentrifugation at 247,600g, 4°C for 21 h (Beckman SW41 rotor). Two major bands (B8 and B9) of spinach PSII–LHCII supercomplexes, corresponding to the *Arabidopsis* B8 and B9 bands¹⁶, were obtained. For electron microscopy studies, the B9 band with a Chl *a/b* ratio of around 2.9–3.1 was collected and concentrated to 3 mg ml^{−1} (in chlorophyll) by using a 100-kDa cutoff concentrator during centrifugation.

The B9 sample was characterized through absorption and fluorescence spectra measurements, SDS–polyacrylamide gel electrophoresis (PAGE), high-performance liquid chromatography (HPLC) and oxygen-evolving activity assay. The absorption spectra were measured at room temperature with a Hitachi U3010 spectrophotometer. The fluorescence spectra were recorded with a Hitachi F7000 fluorescence spectrophotometer at room temperature under 436, 473 and 500 nm excitations. The absorption and fluorescence spectra indicate that a significant amount of Chl *b* (exclusively in the peripheral antenna domain) is present in the sample (Extended Data Fig. 1c, d). For protein composition analysis, 10–18% gradient Tris–Tricine SDS–PAGE was performed in a vertical protein gel electrophoresis system (Protein II, Bio-Rad) according to the protocol described in an earlier report⁴⁸. Four major core subunits (PsbA, B, C and D), three large extrinsic proteins (PsbO, P and Q), Lhcb1/2/4/5 and several small subunits were detected by gel electrophoresis in the denatured B9 sample (Extended Data Fig. 1b). Pigment composition was analysed through HPLC as previously described⁴⁹ with slight modifications. The B9 sample was treated with 80% (v/v) acetone to extract pigments from the supercomplexes and then injected into a C-18 reversed-phase column (Alltech Allsphere ODS-2) in a Hitachi L2130 separation module equipped with a Hitachi L2450 diode array detector. Individual pigments were identified by the absorption spectrum of each elution peak. Pigment analysis indicated that the sample contained Chl *a*, Chl *b*, β -carotene, lutein, neoxanthin and violaxanthin (Extended Data Fig. 1e). The oxygen-evolving activity assay was performed with Chlorolab-2 oxygen electrode system. Grana membranes and B9 samples were diluted into 10 µg ml^{−1} (in chlorophyll) in 2 M betaine, 10 mM NaHCO₃, 10 mM NaCl, 25 mM CaCl₂, 25 mM MES–NaOH (pH 6.5), 0.01% α -DDM. O₂ production was measured at 25°C using 3,773 µmol photons per m² per s white light. The assay was supplied with 0.5 mM 2,5-dichloro-*p*-benzoquinone (DCBQ) as electron acceptors. The supercomplex sample exhibited oxygen-evolving activity at 75 ± 3 µmol O₂ per mg (Chl) per h, comparable to a similar sample prepared from *Arabidopsis* previously¹⁶.

Electron microscopy. Approximately 3.0-µl aliquots of 3 mg ml^{−1} PSII–LHCII supercomplex sample (B9) were applied to glow-discharged GIG holey carbon grids (1.0 µm hole size, 400 mesh). The grid was flash-frozen in liquid ethane at around 100 K using a semi-automatic plunge device (FEI vitrobot IV) with a blotting time of 3 s and blotting force of level 2 at 100% humidity, 16°C. Sample screening was performed on Talos F200C 200-kV electron microscope equipped with a 4 K × 4 K Ceta camera (FEI). The images used for structure determination were collected on a direct electron device (FEI Falcon III) using integrating mode in a 300-kV FEI Titan Krios electron microscope. A total of 1,774 micrographs were recorded at a calibrated magnification of 103,704 yielding a pixel size of 1.35 Å on the detector (detector pixel size: 14 µm), with a dose rate of approximately 25 e[−] Å^{−2} s^{−1} and a defocus range between 0.8 and 2.0 µm. Each exposure of 2 s was dose-fractionated into 32 movie frames.

Data processing, classification and reconstruction. A small data set was collected on an FEI Talos electron microscope with a Ceta camera. Reference-free 2D classification produced several distinguished classes in which some of the class-averaged images could be recognized as side views of the PSII–LHCII supercomplex^{15,16}. Assuming that the two side views with the longest and shortest dimensions were perpendicular to each other, an initial model of the complex was made by using these two side views. The initial model, low-pass filtered to 60 Å, was refined with the whole data set. The refinement yielded an 18 Å-resolution map that was similar to the previous result¹⁵. This reconstruction map was rescaled and used as an initial model for the refinement with the high-resolution data set collected on the Titan Krios.

For the data set collected on the Titan Krios, the beam-induced motion of the whole micrograph with 32 movie frames was corrected by MOTIONCORR⁵⁰. After alignment, an averaged image of 32 frames was used to determine the defocus value and the parameters of astigmatism by program CTFFIND3⁵¹. A subset of around 4,000 particles from about 50 micrographs was first semi-automatically picked using the program e2boxer.py (ref. 52). The images of the subset were classified using reference-free 2D classification and eight of the class-averaged images were selected as templates for an automatic particle-picking procedure in program RELION⁵³ to process the whole data set. The 223,927 particles picked by the program were manually screened to remove those from images with overlapped particles and other bad particles. A total of 192,071 particles were kept for further data processing. These remaining particle images were 2D classified and 182,586 images in good classes were subjected to 3D classification without imposing any symmetry. The class of C₂S₂-type supercomplexes had 143,003 particle images, while the other class of C₂S-type supercomplexes had 39,583 images (excluded from further refinement). After another round of 3D classification, 109,042 images of the C₂S₂ supercomplex were kept for 3D refinement with two-fold symmetry imposed and used to produce a 3D reconstruction map with a nominal resolution of 3.5 Å. The resolution of the reconstruction was further improved to 3.2 Å by local motion correction and particle polishing processes.

Model building and refinement. For model building, crystal structures of *Thermosynechococcus vulcanus* PSII (TvPSII, PDB codes: 3WU2 and 4IL6), Spinach LHCII (PDB code: 1RWT), CP29 (PDB code: 3PL9), PsbP (PDB code: 4RTI) and PsbQ (PDB code: 1VYK) were first manually fitted into the 3.2 Å cryo-EM map in UCSF Chimera⁵⁴ or COOT⁵⁵, and then manually adjusted in COOT. The densities of PsbP and PsbQ in the cryo-EM map are weaker (but clearly distinguishable) than the membrane-intrinsic core subunits (D1, D2, CP43 and CP47), indicating that they may have relatively low occupancy or high mobility. The amino acid sequences of the TvPSII structural model were mutated to its counterparts in *Spinacia oleracea*. The PsbU, PsbV and PsbYcf12 subunits, which are present in TvPSII but absent from higher plants, were deleted during model building. The N-terminal region of CP29 was built manually and based on the well-defined continuous electron density of its main chain, and the sequence was registered according to the bulky side-chain densities in this region. Similarly, *de novo* model building was performed on PsbW and PsbTn. The atomic model of CP26 was mutated from a LHCII monomer and refined according to the cryo-EM map. Among the 15 known low-molecular-mass constituent subunits (14 intrinsic subunits and 1 extrinsic subunit) of plant PSII, 13 have been located in our cryo-EM map of the PSII–LHCII supercomplex. The two unidentified subunits are PsbR and PsbY. PsbR is a 10-kDa protein that is involved in binding PsbP²² and is essential for optimal oxygen-evolving activity of PSII (ref. 56). PsbY was located near PsbE and PsbF in the Sr-substituted TvPSII (ref. 57). In the spinach PSII–LHCII supercomplex, no strong protein density corresponding to PsbR or PsbY was observed, indicating that they might be lost during purification. PsbS is essential for photoprotection through non-photochemical quenching^{58,59}. Although the B9 sample used for the cryo-EM study contained PsbS protein, it was not observed in the electron density map of the supercomplex, probably owing to its nonspecific association with the supercomplex, as explained previously¹⁶.

The structure model of the spinach PSII–LHCII supercomplex was first refined in real space against the cryo-EM map by Phenix 1.9 (ref. 60) with geometry and secondary structure restraints. During real space refinement, the distances between the central magnesium ions of chlorophyll molecules and the coordinating ligands were restrained according to the values obtained from the high-resolution crystal structures. In addition, refinement in reciprocal space was performed in REFMAC^{61,62} with stereo-chemical and homology-derived restraints using modified scripts of the program adapted for the cryo-EM map. Automatic real-space and reciprocal-space refinements followed by manual correction in COOT were carried out iteratively until there were no more improvements in both *R* factor and geometry parameters. The statistics for data collection and structure refinement are summarized in Extended Data Fig. 2c.

48. Schagger, H. Tricine-SDS-PAGE. *Nature Protocols* **1**, 16–22 (2006).

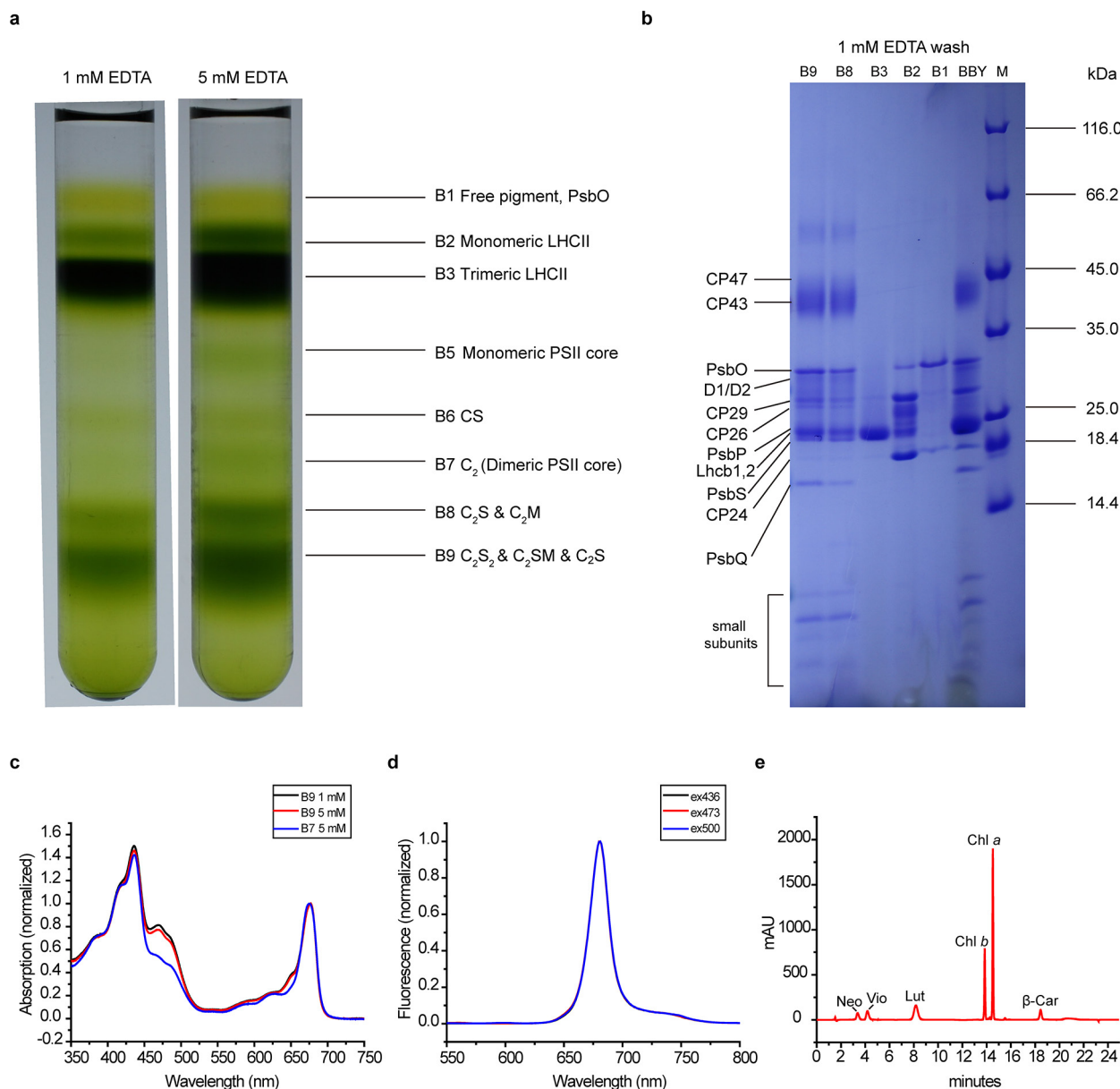
49. Färber, A., Young, A. J., Ruban, A. V., Horton, P. & Jahns, P. Dynamics of xanthophyll-cycle activity in different antenna subcomplexes in the photosynthetic membranes of higher plants (the relationship between zeaxanthin conversion and nonphotochemical fluorescence quenching). *Plant Physiol.* **115**, 1609–1618 (1997).

50. Li, X. *et al.* Electron counting and beam-induced motion correction enable near-atomic-resolution single-particle cryo-EM. *Nature Methods* **10**, 584–590 (2013).

51. Mindell, J. A. & Grigorieff, N. Accurate determination of local defocus and specimen tilt in electron microscopy. *J. Struct. Biol.* **142**, 334–347 (2003).

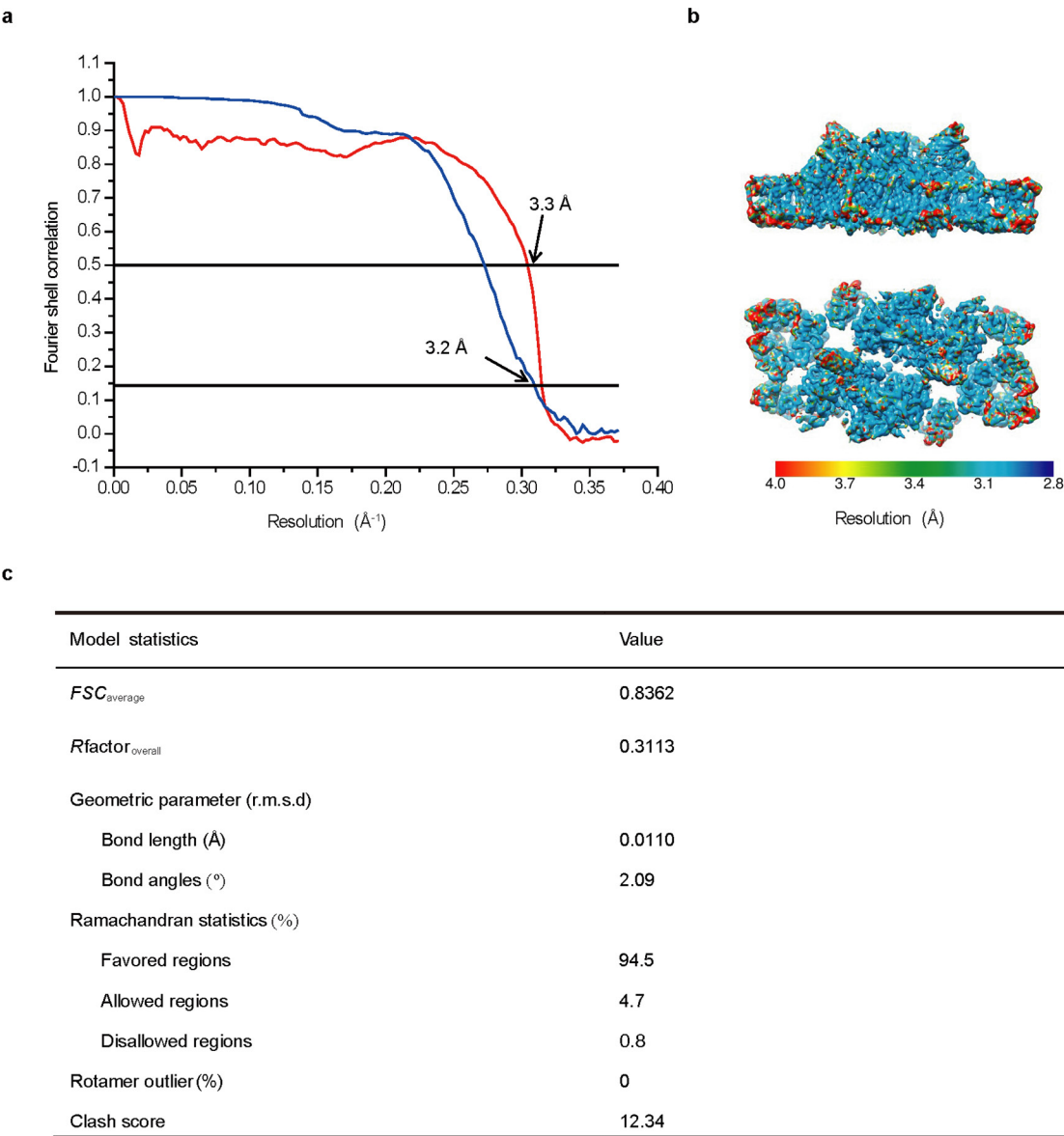
52. Tang, G. *et al.* EMAN2: an extensible image processing suite for electron microscopy. *J. Struct. Biol.* **157**, 38–46 (2007).

53. Scheres, S. H. W. RELION: Implementation of a Bayesian approach to cryo-EM structure determination. *J. Struct. Biol.* **180**, 519–530 (2012).
54. Pettersen, E. F. *et al.* UCSF Chimera—a visualization system for exploratory research and analysis. *J. Comput. Chem.* **25**, 1605–1612 (2004).
55. Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. Features and development of Coot. *Acta Crystallogr. D* **66**, 486–501 (2010).
56. Allahverdiyeva, Y. *et al.* Insights into the function of PsbR protein in *Arabidopsis thaliana*. *Biochim. Biophys. Acta* **1767**, 677–685 (2007).
57. Koua, F. H. M., Umena, Y., Kawakami, K. & Shen, J.-R. Structure of Sr-substituted photosystem II at 2.1 Å resolution and its implications in the mechanism of water oxidation. *Proc. Natl Acad. Sci. USA* **110**, 3889–3894 (2013).
58. Li, X. P. *et al.* A pigment-binding protein essential for regulation of photosynthetic light harvesting. *Nature* **403**, 391–395 (2000).
59. Fan, M. *et al.* Crystal structures of the PsbS protein essential for photoprotection in plants. *Nature Struct. Mol. Biol.* **22**, 729–735 (2015).
60. Adams, P. D. *et al.* PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr. D* **66**, 213–221 (2010).
61. Murshudov, G. N. *et al.* REFMAC5 for the refinement of macromolecular crystal structures. *Acta Crystallogr. D* **67**, 355–367 (2011).
62. Winn, M. D. *et al.* Overview of the CCP4 suite and current developments. *Acta Crystallogr. D* **67**, 235–242 (2011).
63. Green, B. R. & Durnford, D. G. The chlorophyll-carotenoid proteins of oxygenic photosynthesis. *Annu. Rev. Plant Physiol. Plant Mol. Biol.* **47**, 685–714 (1996).
64. Das, S. K. & Frank, H. A. Pigment compositions, spectral properties, and energy transfer efficiencies between the xanthophylls and chlorophylls in the major and minor pigment–protein complexes of photosystem II. *Biochemistry* **41**, 13087–13095 (2002).
65. Pascal, A. *et al.* Spectroscopic characterization of the spinach Lhcb4 protein (CP29), a minor light-harvesting complex of photosystem II. *Eur. J. Biochem.* **262**, 817–823 (1999).
66. van Amerongen, H. *et al.* Spectroscopic characterization of CP26, a chlorophyll *ab* binding protein of the higher plant Photosystem II complex. *Biochim. Biophys. Acta* **1188**, 227–234 (1994).



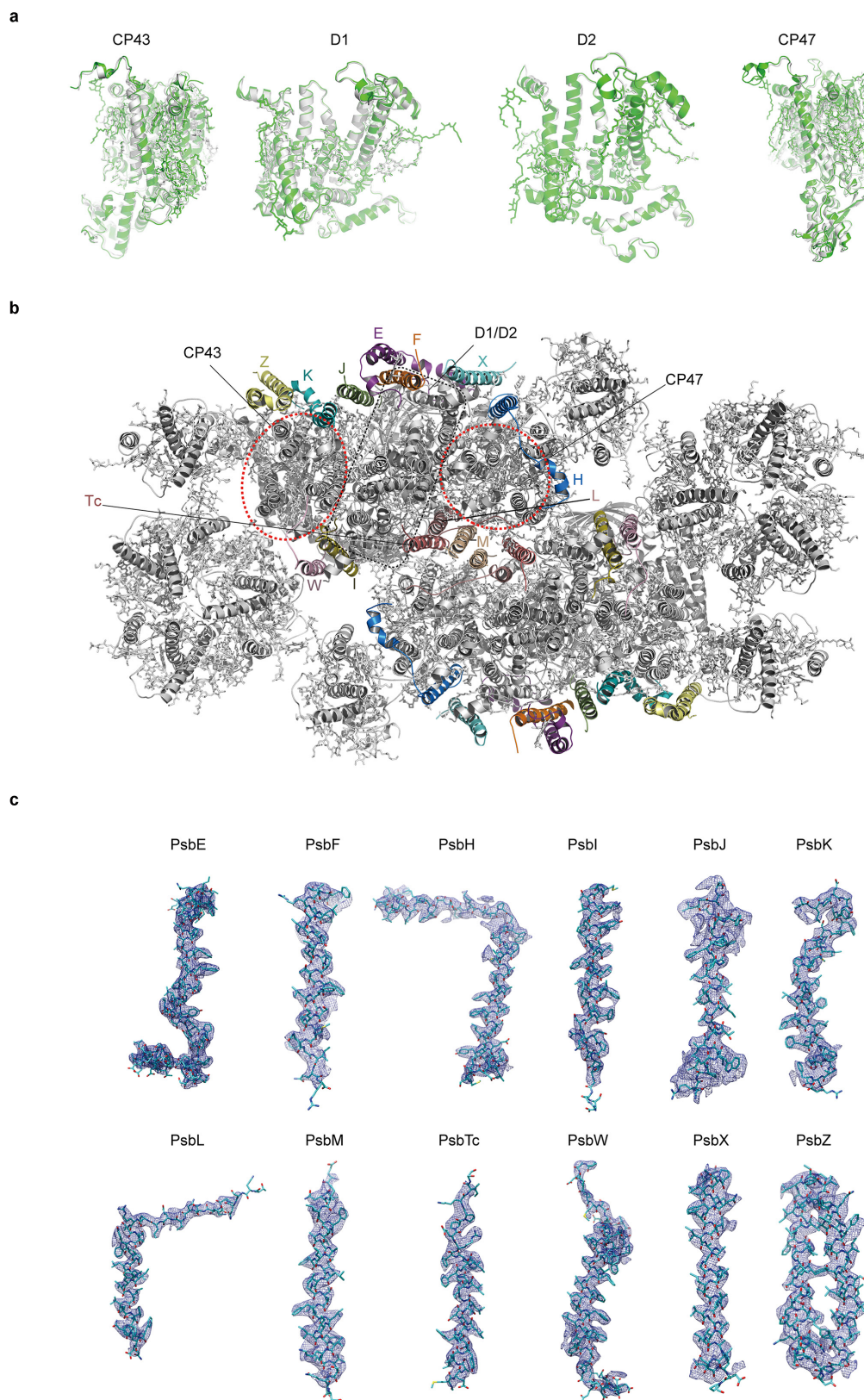
Extended Data Figure 1 | Purification and characterization of the spinach PSII-LHCII supercomplex. **a**, Sucrose gradient of solubilized grana membranes. The membrane preparations were first washed with 1 mM (left) or 5 mM (right) EDTA before being solubilized by α -DDM for further purification through sucrose-gradient ultracentrifugation. The content of each band is indicated based on the absorption spectrum and SDS-PAGE results, and by comparing to previously published data. The B9 fraction obtained from the grana membrane washed with 1 mM EDTA was used for cryo-EM. Note that the grana membranes washed with 1 mM EDTA yielded less B5, B6 and B7 than the sample treated with 5 mM EDTA. **b**, SDS-PAGE analysis of the sucrose gradient fractions. The protein composition of each Coomassie band was indicated based on the mass spectrometry and proteomics data analysis. For gel source data, see Supplementary Fig. 1. **c**, Room-temperature absorption spectrum of B9 sample used for cryo-EM. Its spectrum (B9 1 mM) is compared to those of

B7 (dimeric PSII core without LHCII attached; B7 5 mM) and B9 samples (B9 5 mM) fractionated from grana washed with 5 mM EDTA. Note that B9 from grana membranes washed with 1 mM EDTA showed higher peaks at 470 and 650 nm, indicating that this fraction contains higher Chl *b* content (from LHCII) than the other two. The spectra are normalized to the maximum in the red region. **d**, Fluorescence emission spectra of B9 sample measured at room temperature. The maximum emissions were at 681 nm (upon excitation of Chl *a* at 436 nm), 680 nm (upon excitation of Chl *b* at 473 nm) and 681 nm (upon excitation of carotenoids at 500 nm). Overlapping of these three spectra suggests that nearly all pigments in the B9 sample are well coupled and no free pigments are present. **e**, Pigment content analysis of B9 sample by HPLC. Based on the characteristic absorption spectrum of each peak fraction, the six major pigment peaks separated from the B9 sample are identified as neoxanthin (Neo), violaxanthin (Vio), lutein (Lut), Chl *b*, Chl *a* and β -carotene (β -car).



Extended Data Figure 2 | Evaluation of the resolution of the cryo-EM structure of the spinach PSII-LHCII supercomplex. **a**, Fourier shell correlation (FSC) plots. Blue, gold-standard FSC curve with a value of 0.143 at 3.2 Å resolution; red, FSC curve calculated between the cryo-EM map and the refined structure model of the PSII-LHCII supercomplex. The map-model FSC has a value of 0.5 at 3.3 Å resolution. **b**, Local

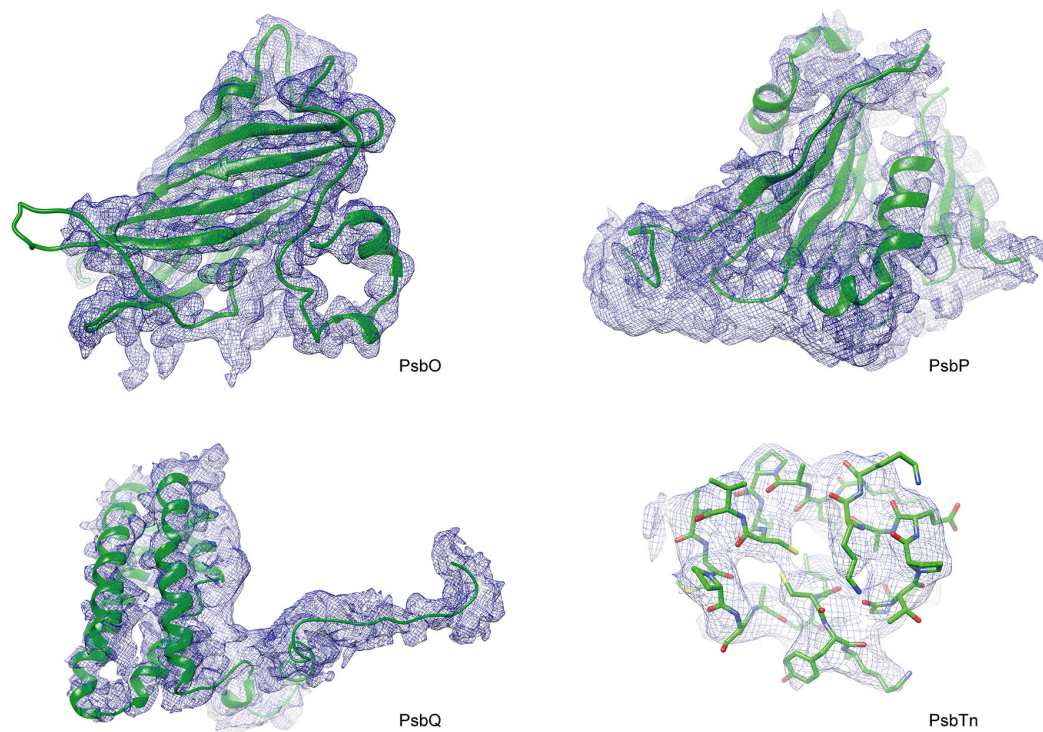
resolutions of the cryo-EM map of the spinach PSII-LHCII supercomplex estimated by Resmap. Top, side view along the membrane plane with the luminal domain facing upwards. Bottom, bottom view from the luminal side and approximately along the membrane normal (or C2 axis). **c**, The statistics of the structural model of the spinach PSII-LHCII supercomplex refined against the 3.2 Å resolution cryo-EM map.



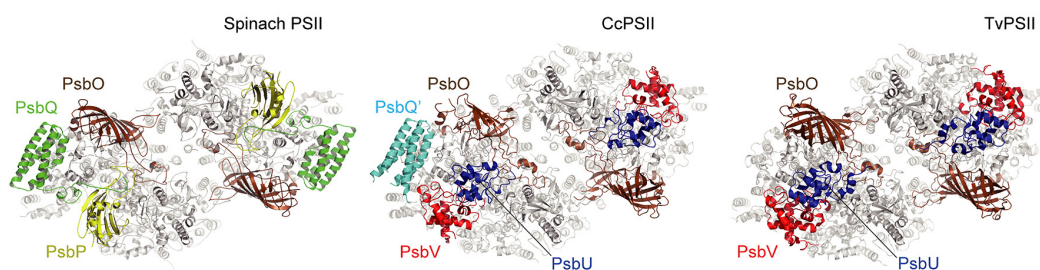
Extended Data Figure 3 | Structures of the large and small intrinsic subunits of spinach PSII. **a**, The four large intrinsic subunits of the spinach PSII core superposed on the corresponding subunits of the TvPSII core. The protein backbones and cofactors are shown as ribbon and stick models, respectively. Silver, spinach PSII core subunits; green, TvPSII core

subunits. **b**, The locations of 12 low-molecular-mass intrinsic subunits in the spinach PSII-LHCII supercomplex. These subunits are coloured and the rest of the supercomplex is grey. **c**, The densities for the low-molecular-mass intrinsic subunits are shown as blue meshes. The corresponding models are shown as cyan sticks.

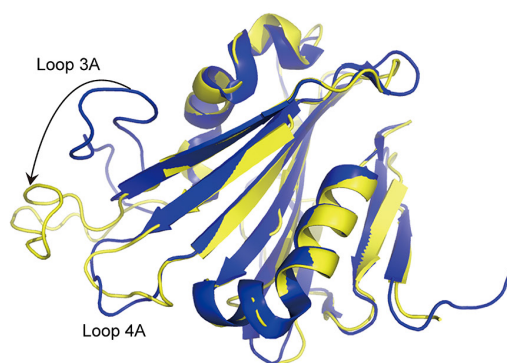
a



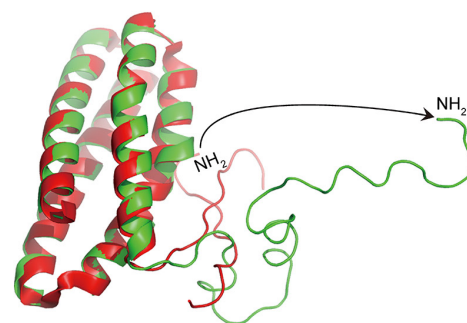
b



c

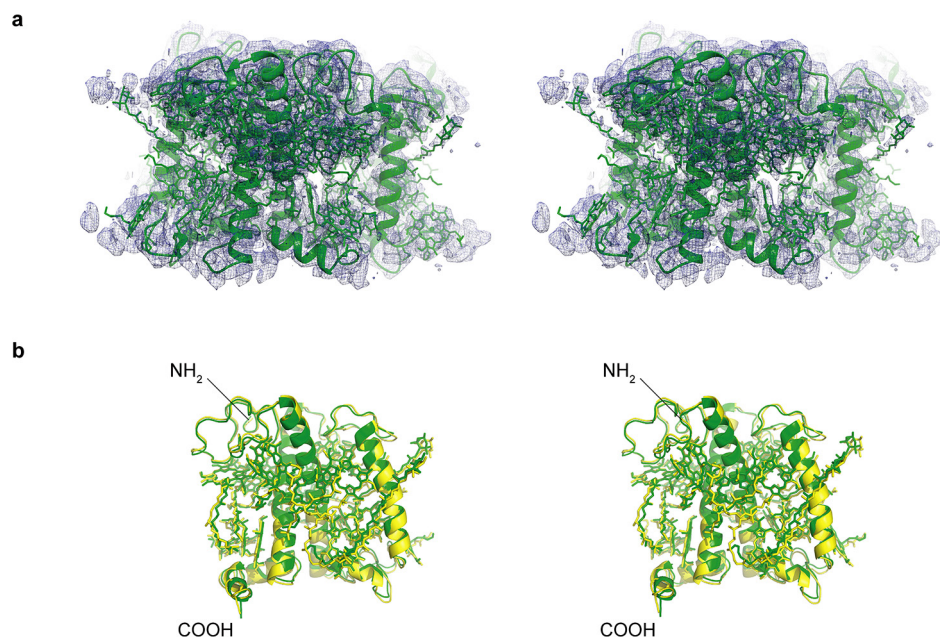


d



Extended Data Figure 4 | Cryo-EM densities and structures of the extrinsic subunits. **a**, Cryo-EM densities of PsbO, PsbP, PsbQ and PsbTn. **b**, The binding sites of spinach PsbP, PsbQ and PsbO compared to those of the extrinsic subunits in CcPSII and TvPSII. The spinach PSII core is shown at an angle identical to that of the CcPSII/TvPSII core. PDB codes: 4YUU (CcPSII); 3WU2 (TvPSII). **c**, Superposition of PsbP bound in the supercomplex with the isolated PsbP. Colour code: yellow, PsbP in the supercomplex; blue, isolated PsbP (PDB code: 4RTI). Loop 3A and Loop 4A

indicate the loop regions between Lys90 and Ala111 and between Arg134 and Gly142, respectively. Note the conformational change in Loop 3A (arrow) when PsbP binds to the PSII core. **d**, Structure of PsbQ bound in the supercomplex superposed with the isolated PsbQ. Green, PsbQ in the supercomplex; red, isolated PsbQ (PDB code: 1VYK). Note the conformational change in the elongated N-terminal region from a folded state to an extended form (arrow) when PsbQ binds to the PSII core.

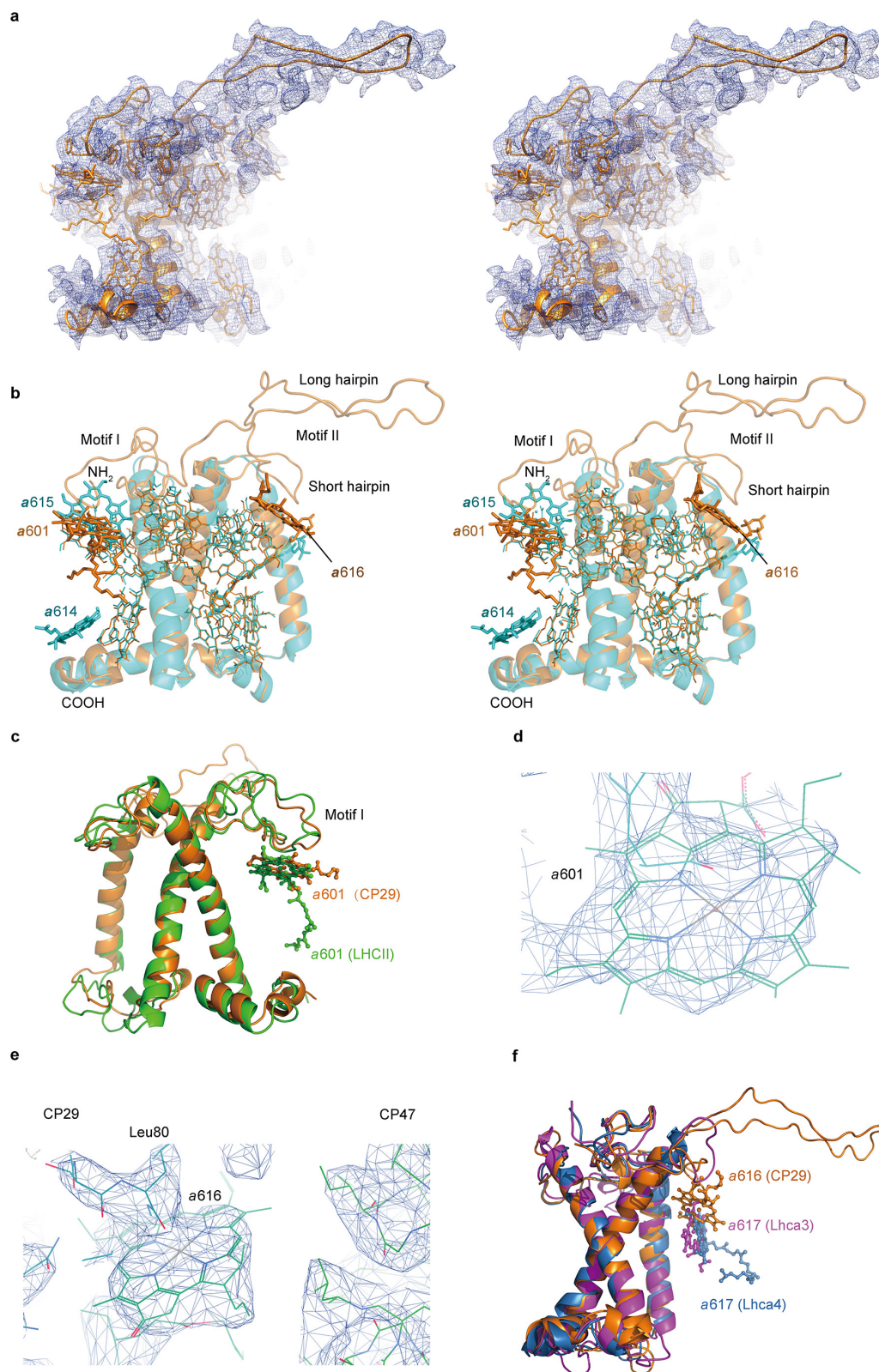


Extended Data Figure 5 | Cryo-EM density and structure of LHCII.

a, Cryo-EM densities of the LHCII trimer in the supercomplex. Stereo pairs are shown and the view is along the membrane plane.

b, Superposition of the cryo-EM structure of an LHCII monomer with

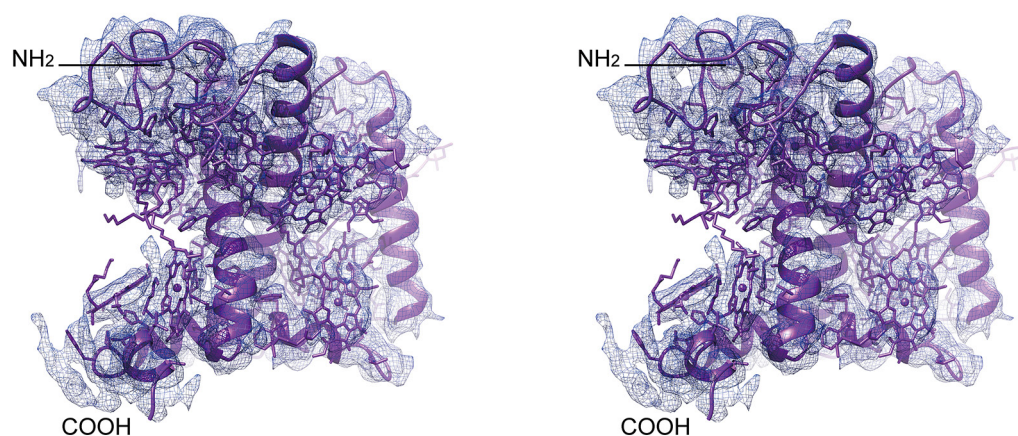
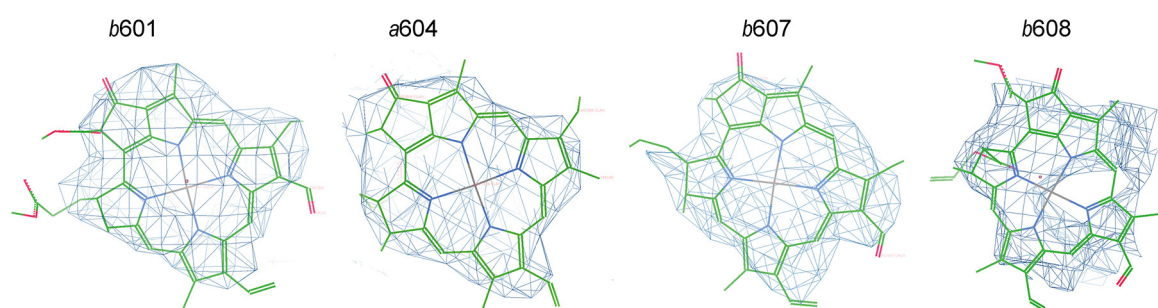
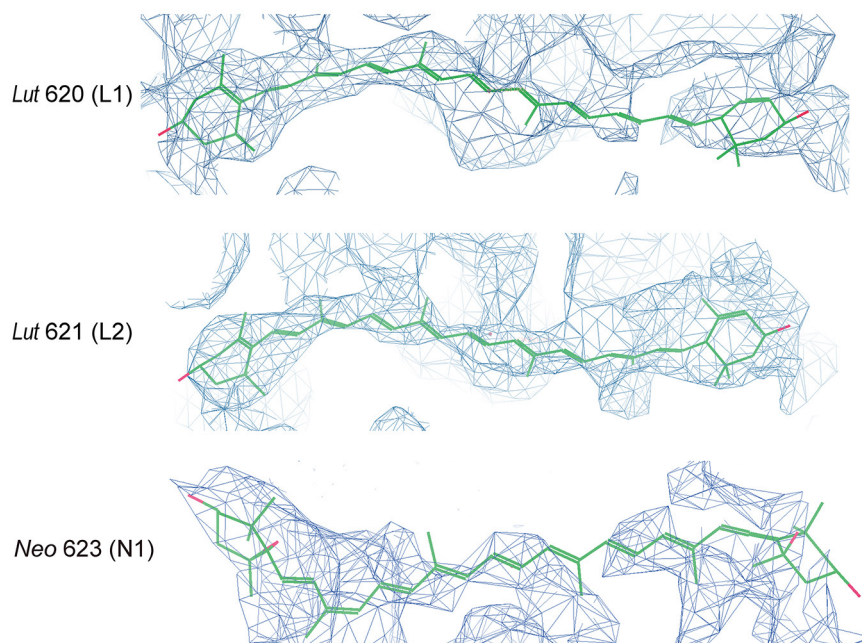
the previous crystal structure (PDB code: 1RWT). The protein backbone is shown as ribbon diagrams and the cofactors are displayed as stick models. Green, cryo-EM structure; yellow, crystal structure.



Extended Data Figure 6 | Cryo-EM density and structure of CP29.

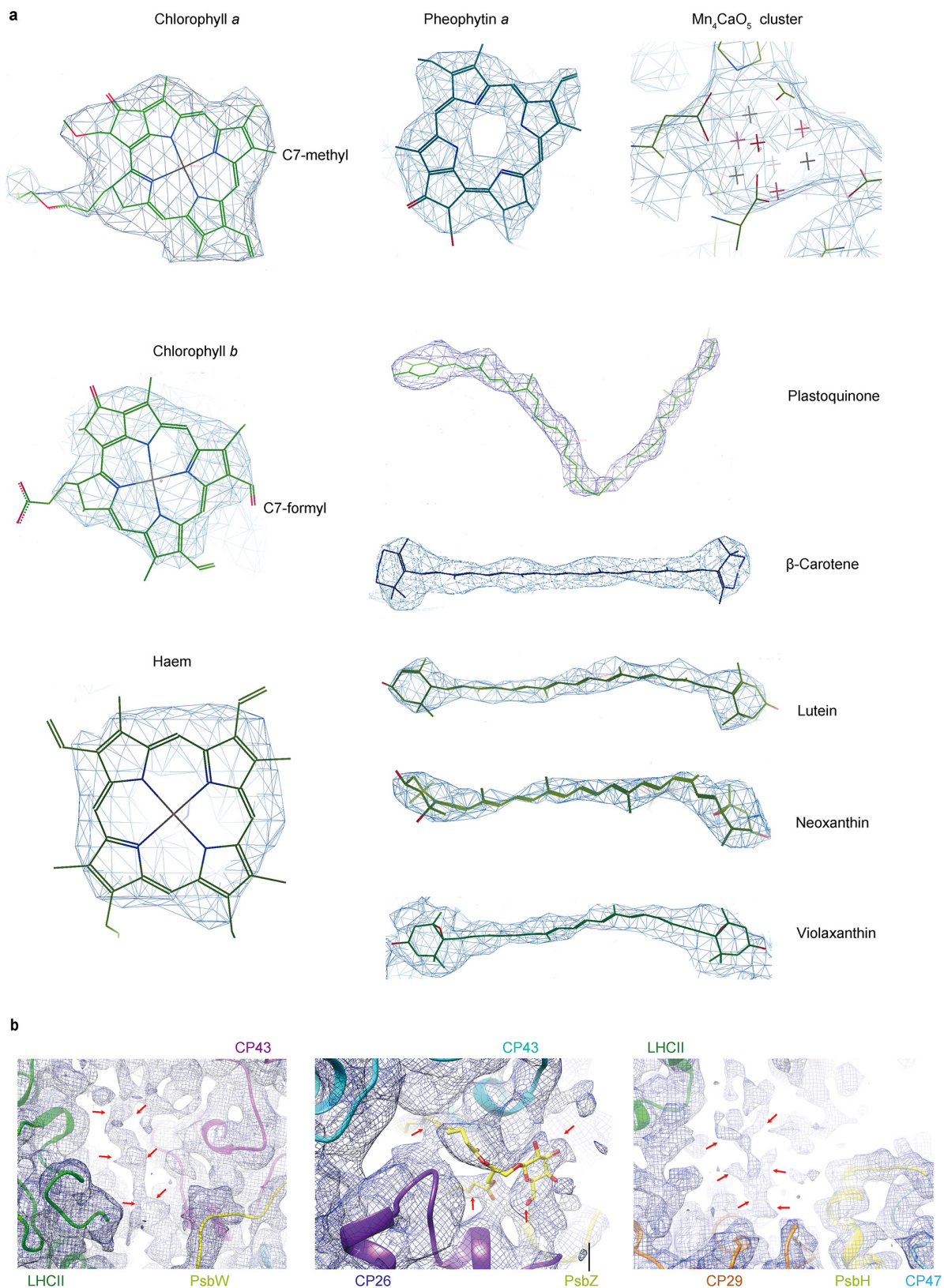
a, Stereo image of the cryo-EM density of CP29 bound in the PSII-LHCII supercomplex. **b**, Superposition of cryo-EM structure of full-length CP29 with the previous crystal structure. Note: Chls *a*601 and *a*616 are newly observed in the cryo-EM structure of CP29. Chl *a*601 might account for the electron density of Chl *a*615 observed in the crystal structure of spinach CP29 (ref. 14). Compared to Chl *a*601, *a*615 is much closer to *a*611 owing to the loss of the N-terminal domain caused by proteolysis. Chl *b*614 is a peripheral chlorophyll found in the crystal structure, but is probably lost during purification and therefore not observed in the

cryo-EM structure. Orange, cryo-EM structure; cyan, crystal structure. **c**, Superposition of CP29 (orange) with the structure of an LHCII monomer (green). For the cofactors, only Chl *a*601 is shown; the others are omitted for clarity. **d**, Cryo-EM density of Chl *a*601 in CP29. **e**, Cryo-EM density of Chl *a*616 at the interface between CP29 and CP47. **f**, Superposition of Lhca3 and Lhca4 structures with that of CP29 in the PSII-LHCII supercomplex. Chl *a*616 (CP29) and *a*617 (Lhca3/4) molecules are shown as stick models; the other cofactors are omitted for clarity. Orange, CP29; magenta, Lhca3; blue, Lhca4. PDB codes: 4XK8, Lhca3 and Lhca4 from the PSI-LHCI supercomplex; 1RWT, LHCII.

a**b****c**

Extended Data Figure 7 | Cryo-EM density and structure of CP26 bound in the PSII-LHCII supercomplex. **a**, Stereo images of the density and overall structure of CP26. The density is shown as grey meshes and the model is in purple. The protein backbone is shown as a ribbon model; the cofactors are presented as stick models. **b**, The densities for Chl b601,

Chl a604, Chl b607 and Chl b608 in CP26. These four chlorophylls were not predicted in the previous work, but are clearly present in the structure. **c**, The density for three carotenoids in CP26. Note that the density for the epoxidized head group of neoxanthin is clearly visible, while the rest of it is fairly weak (presumably owing to low occupancy or high flexibility).



Extended Data Figure 8 | Cryo-EM densities of various cofactors bound in the spinach PSII-LHCII supercomplex. a, The densities of chlorophylls, carotenoids, Mn_4CaO_5 and plastoquinone molecules. **b,** The potential lipid densities at the interfacial regions between adjacent antenna complexes. The interfaces between LHCII and PsbW, CP26 and CP43,

and LHCII and CP29 are shown from left to right. Red arrows indicate the positions of potential lipid densities. The cryo-EM densities are displayed as grey meshes and the atomic models for interpretation of the densities are shown as sticks and bullets.

Extended Data Table 1 | Cofactors located within each monomer of the spinach PSII-LHCII supercomplex

Protein	chlorophyll	carotenoid	haem	lipid	others
D1 (PsbA)	4 Chl <i>a</i> and 2 Pheophytin	1 BCR		2 SQDG 1 MGDG	1 Mn ₄ CaO ₅ cluster, <1 plastoquinone (weak density due to partial occupancy on Q _B site)
D2 (PsbD)	2 Chl <i>a</i>	1 BCR		3 PG 1 MGDG	1 plastoquinone (strong density on Q _A site)
CP47 (PsbB)	16 Chl <i>a</i>	3 BCR		1 SQDG 1 MGDG	
CP43 (PsbC)	13 Chl <i>a</i>	3 BCR		3 DGDG 1 MGDG	
PsbH		1 BCR		1 DGDG	
PsbK		1 BCR			
PsbL				1 PG	
PsbZ				1 MGDG	
Cyt <i>b</i> 559 (PsbE & F)	-	-	1		
LHCII trimer	42 (24 Chl <i>a</i> and 18 Chl <i>b</i>)	12 (6 Lut, 3 Vio, 3 Neo)		3 PG	
CP29	13 (10 Chl <i>a</i> and 3 Chl <i>b</i>)	3 (1 Lut, 1 Vio, 1 Neo)		1 PG	
CP26	13 (9 Chl <i>a</i> and 4 Chl <i>b</i>)	3 (2 Lut, 1 Neo)		1 PG	
Total	105	28	1	21	2-3

BCR, β -carotene; DGDG, digalactosyldiacyl glycerol; Lut, lutein; MGDG, monogalactosyldiacyl glycerol; Neo, neoxanthin; PG, phosphatidyl glycerol; SQDG, sulfoquinovosyldiacyl glycerol; Vio, violaxanthin.

Extended Data Table 2 | Pigment binding sites of spinach LHCII, CP29 and CP26 in the PSII-LHCII supercomplex

Peripheral antenna complexes	LHCII	CP29	CP26
Chlorophylls			
601	Chl <i>b</i> (Tyr24) [†]	Chl <i>a</i> (Trp14) [†]	Chl <i>b</i> (Phe34) [‡]
602	Chl <i>a</i> (Glu65)	Chl <i>a</i> (Glu96)	Chl <i>a</i> (Glu78)
603	Chl <i>a</i> (His68)	Chl <i>a</i> (His99)	Chl <i>a</i> (His81)
604	Chl <i>a</i> (H ₂ O)	Chl <i>a</i> (H ₂ O)	Chl <i>a</i> (putative H ₂ O) [‡]
605	Chl <i>b</i> (Val119)[Gln122, Ser123] [§]	-	-
606	Chl <i>b</i> (H ₂ O)	Chl <i>b</i> (H ₂ O)	Chl <i>b</i> (putative H ₂ O)
	[H ₂ O]	[putative H ₂ O]	[putative H ₂ O]
607	Chl <i>b</i> (H ₂ O)[Gln131]	Chl <i>b</i> (H ₂ O)[Glu151]	Chl <i>b</i> (putative H ₂ O)
			[Glu142] [‡]
608	Chl <i>b</i> (H ₂ O)[Leu148]	Chl <i>b</i> (H ₂ O)[Gln161]	Chl <i>b</i> (H ₂ O) [‡]
609	Chl <i>b</i> (Glu139)[Gln131]	Chl <i>a</i> (Glu159)	Chl <i>a</i> (may also accept Chl <i>b</i>) (Glu150)
610	Chl <i>a</i> (Glu180)	Chl <i>a</i> (may also accept Chl <i>b</i>) (Glu197)	Chl <i>a</i> (Glu189)
611	Chl <i>a</i> (PG)	Chl <i>a</i> (PG)	Chl <i>a</i> (PG)
612	Chl <i>a</i> (Asn183)	Chl <i>a</i> (His200)	Chl <i>a</i> (Asn192)
613	Chl <i>a</i> (Gln197)	Chl <i>a</i> (Gln214)	Chl <i>a</i> (Gln206)
614	Chl <i>a</i> (His212)		Chl <i>a</i> (His221)
615			-
616		Chl <i>a</i> (Leu80) [‡]	-
Chl <i>a/b</i> ratio (structural model)	1.33	3.3 (or 2.5 if Chl <i>b</i> 614 is present)	2.3 (may be lower if 609 is an Chl <i>a/b</i> mixed site)
Chl <i>a/b</i> ratio (biochemical analyses) [¶]	1.3-1.4	2.5-3.0	2.1-3.3
Carotenoids			
L1	lutein	lutein	lutein
L2	lutein	violaxanthin	lutein
			(may also accept violaxanthin)
N1	neoxanthin	neoxanthin	neoxanthin
V1	violaxanthin	-	-

The local resolution of our cryo-EM map has an uneven distribution, as shown in Extended Data Fig. 2b. The core region has a relatively higher resolution (at 3.0–3.5 Å) than in the regions of peripheral antenna system (3.2–4.0 Å), sufficient to identify the number of pigment molecules bound to each antenna complexes and locate their positions. The identities of chlorophylls (Chl *a* or Chl *b*) and carotenoids (lutein, neoxanthin or violaxanthin) are assigned mainly by referring to the information obtained from previous work on the high-resolution crystal structures of spinach LHCII¹² and CP29 (ref. 14) and the functional architecture of CP26 (ref. 23).

*Central ligands of chlorophylls coordinating the Mg atoms are shown in parentheses.

†As the N-terminal region of CP29 is intact in the cryo-EM structure, its 601 site is occupied by a chlorophyll (tentatively assigned as Chl *a*) coordinated by Trp14 (corresponding to Tyr24 in LHCII). Owing to proteolysis at the N-terminal region, the chlorophyll at the 601 site in the previous crystal structure of CP29 might have shifted to the nearby 615 site, sharing the same ligand with Chl *a*611.

‡Newly identified chlorophyll-binding site in CP29 or CP26.

§The hydrogen bond donors of the C7-formyl group of Chl *b* molecules are shown in square brackets.

||These sites in the previous crystal structure of CP29 were occupied by a Chl *b* (614) and Chl *a* (615) coordinated by His229 (614) and glycerol-3-phosphate, respectively. They are not observed in the cryo-EM structure reported here. Chl *b*614 is located at a peripheral site in contact with detergent and might be lost during purification, leading to a vacant site without chlorophyll bound.

¶These data were extracted and summarized from previous publications^{14,23,63–66}.

Regulation of black-hole accretion by a disk wind during a violent outburst of V404 Cygni

T. Muñoz-Darias^{1,2}, J. Casares^{1,2,3}, D. Mata Sánchez^{1,2}, R. P. Fender³, M. Armas Padilla^{1,2,4}, M. Linares^{1,2,5}, G. Ponti⁶, P. A. Charles^{3,7}, K. P. Mooley³ & J. Rodríguez⁸

Accretion of matter onto black holes is universally associated with strong radiative feedback¹ and powerful outflows². In particular, black-hole transients³ have outflows whose properties⁴ are strongly coupled to those of the accretion flow. This includes X-ray winds of ionized material, expelled from the accretion disk encircling the black hole, and collimated radio jets^{5,6}. Very recently, a distinct optical variability pattern has been reported in the transient stellar-mass black hole V404 Cygni, and interpreted as disrupted mass flow into the inner regions of its large accretion disk⁷. Here we report observations of a sustained outer accretion disk wind in V404 Cyg, which is unlike any seen hitherto. We find that the outflowing wind is neutral, has a large covering factor, expands at one per cent of the speed of light and triggers a nebular phase once accretion drops sharply and the ejecta become optically thin. The large expelled mass ($>10^{-8}$ solar masses) indicates that the outburst was prematurely ended when a sizeable fraction of the outer disk was depleted by the wind, detaching the inner regions from the rest of the disk. The luminous, but brief, accretion phases shown by transients with large accretion disks² imply that this outflow is probably a fundamental ingredient in regulating mass accretion onto black holes.

The X-ray binary V404 Cyg (GS 2023+338) is a confirmed stellar-mass black hole⁸ with a precisely determined distance from Earth of 2.4 kpc (ref. 9). After 25 years of quiescence, NASA's Swift mission detected renewed activity on 15 June 2015¹⁰, initiating a two-week period of intensely violently variable emission across all wavelengths^{11,12}. Our high signal-to-noise optical spectra covering the entire X-ray/radio-active phase (~ 15 days) show that, contemporaneously with radio jet emission, continuous ejections of neutral material at $\sim 0.01c$, where c is the speed of light in a vacuum, are present from low-level accretion phases ($<1\%$ of the Eddington luminosity L_{Edd}) to the X-ray peak (Methods; Fig. 1, Extended Data Fig. 1). These are observed in hydrogen (Balmer) and helium (He I) emission lines as deep P Cyg profiles throughout the outburst¹³, and extremely broad wings once the X-ray and radio fluxes decay. P Cyg profiles result from resonant scattering in an expanding outflow with a spherical geometry or at least sustaining a large solid angle^{14,15} (Methods). Of a dozen transitions showing this feature, the deepest are seen in the He I, $\lambda = 5,876 \text{ \AA}$ emission line, which is used as a reference for this study (see Extended Data Fig. 2).

The strongest P Cyg profiles are witnessed during days 1 to 6 (Fig. 1 and Fig. 2 for the evolution of the profiles during day 2; see Methods), when the X-ray luminosity is typically 1,000 times fainter than the $\sim L_{\text{Edd}}$ flares displayed later in the outburst^{7,11} (Extended Data Fig. 1). Blue-shifted absorptions are as deep as 30% below the continuum level and we measure terminal velocities in the range $V_T = 1,500\text{--}3,000 \text{ km s}^{-1}$ (Figs 1 and 2, Extended Data Figs 2 and 3). Symmetric

red-shifted (that is, positive velocity) outflow emission, completely detached from the accretion disk line component, is sometimes evident (see Fig. 2 from minute 60 onwards).

To trace the ionization state of the outer disk, we computed the line flux ratio $I_{\text{ratio}} = \text{He II } (\lambda = 4,686 \text{ \AA})/\text{H}\beta$ and obtain $I_{\text{ratio}} < 0.5$ when P Cyg absorptions are deepest (Extended Data Fig. 1; Methods). Through the outburst, both the X-ray and optical emission are characterized by the presence of short and long flaring activity^{7,11}. During these flaring episodes, I_{ratio} increases while the P Cyg profiles become weaker, subsequently recovering their pre-flare strength when the X-ray flux and I_{ratio} drop (Fig. 2). This indicates that the detection of P Cyg absorptions is driven by ionization effects. Indeed, on days 7 to 10 much shallower absorptions (only $\sim 2\%$ below the continuum level) are witnessed as the system enters the brightest phase of the outburst and I_{ratio} becomes always larger than unity (Fig. 1 and Extended Data Fig. 1). Furthermore, we note that the H α profile is very asymmetric during the whole outburst, providing a further indication of the ubiquitous presence of wind outflows during our observations (Extended Data Fig. 4; Methods).

The low temperature T that is required to have both neutral hydrogen ($T < 10^4 \text{ K}$) and helium ($T < 3 \times 10^4 \text{ K}$) places the wind-launching radius R_l at the outer accretion disk regardless of the wind-launching mechanism. On the other hand, the low luminosity associated with the deepest P Cyg profiles rules out radiation-pressure winds driven by Thomson scattering. The thermal wind scenario¹⁶, in which V_T roughly corresponds to the escape velocity at R_l , is able to reproduce our observations. Using $V_T = 1,500\text{--}3,000 \text{ km s}^{-1}$, we obtain $R_l = (1.5\text{--}6) \times 10^5 \text{ km}$, which corresponds to disk temperatures in the range $\sim 5,000\text{--}30,000 \text{ K}$ for luminosities within the range $0.001 L_{\text{Edd}}\text{--}0.1 L_{\text{Edd}}$ (Methods). A crude estimation of the mass-loss rate associated with the most conspicuous profiles (day 2) suggests $\dot{M}_{\text{out}} > 10^{-13} M_{\odot} \text{ yr}^{-1}$ (Methods). This lower limit accounts only for neutral matter outflows, but a wind of ionized material could also be launched. A hot wind with V_T up to $\sim 4,000 \text{ km s}^{-1}$ is indeed detected by the only Chandra pointing performed during the outburst¹⁷.

The second signature of the high-velocity wind is a short nebular phase witnessed at the end of the outburst. Following a sharp drop by a factor of $\sim 1,000$ in the X-ray, optical and radio luminosity from the major flares that end the brightest phase of the outburst on days 9 and 10, the Balmer lines became unprecedentedly intense for a black hole, showing equivalent widths up to $\sim 2,000 \text{ \AA}$ (H α ; Extended Data Fig. 1). They sit on extended wings reaching similar velocities to the V_T observed in the P Cyg profiles ($\pm 3,000 \text{ km s}^{-1}$; inset in Fig. 3). A forest of broad emission lines, such as Si II and Fe II, also appears (Fig. 3), while the Balmer decrement (BD; the ratio between the Balmer line fluxes; see Methods) increases up to ~ 6 , as compared to ~ 2.5 observed earlier (Extended Data Figs 1 and 5). High values of BD

¹Instituto de Astrofísica de Canarias, E-38205 La Laguna, Santa Cruz de Tenerife, Spain. ²Departamento de Astrofísica, Universidad de La Laguna, E-38206 La Laguna, Santa Cruz de Tenerife, Spain. ³Department of Physics, Astrophysics, University of Oxford, Denys Wilkinson Building, Keble Road, Oxford OX1 3RH, UK. ⁴Department of Astronomy, Kyoto University, Kyoto 606-8502, Japan. ⁵Institutt for Fysikk, Norges Teknisk-Naturvitenskapelige Universitet (NTNU), Trondheim, Norway. ⁶Max-Planck-Institut für extraterrestrische Physik, Giessenbachstrasse 1, D-85748 Garching bei München, Germany. ⁷School of Physics and Astronomy, University of Southampton, Southampton SO17 1BJ, UK. ⁸Laboratoire Astrophysique Instrumentation Modélisation (AIM), UMR 7158, CEA/CNRS/Université Paris Diderot, CEA DRF/IRFU/SAp, 91191 Gif-sur-Yvette, France.

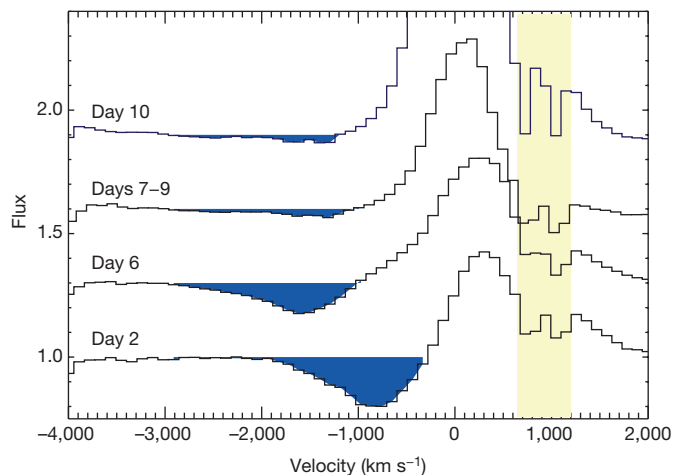


Figure 1 | P Cyg profiles observed during days 2, 6, 7–9 and 10 in He I $\lambda = 5,876$ Å. Normalized spectra are offset by 0, 0.6, 1.2 and 1.8, respectively. Profiles are formed when atomic material approaching the observer at velocity $-V_{\text{out}}$ (with V_{out} the projected outflow velocity) scatters photons with frequency $\nu = \nu_0(1 - V_{\text{out}}/c)$, while receding ejecta moving at velocity V_{out} are being illuminated by the central source. Yellow shading indicates regions contaminated by interstellar absorption. We detect approaching material moving at up to $3,000 \text{ km s}^{-1}$ (blue-filled absorptions). During days 7–9 and day 10 the profiles are very shallow, corresponding to high ionization states (see text). Simultaneously with the blue-shifted absorption we detect red-shifted emission detached from the accretion disk component (see also Fig. 2). This red-shifted emission reached amplitudes similar to that of emission lines produced by approaching material, a feature indicating spherical geometry or at least a large covering factor¹⁴.

are associated with nebularities, as a result of neutral hydrogen self-absorption in relatively low-density conditions¹⁸ (Methods). This behaviour is expected when the outflow cools and expands, becoming optically thin. The symmetric wings indicate a large covering factor for the ejecta. Expanding nova shells are characterized by similar BD values¹⁹ during some stages, as well as exhibiting some of the emission

lines detected here. These emission lines are also found in low-excitation nebularities surrounding outflowing massive stars²⁰, which show similar H α equivalent widths in their final and most violent evolutionary phases²¹. This phase is not witnessed after other strong flares displayed early in the outburst (for example, day 4). However, these events are not followed by a strong drop in flux, such as occurs in the case of the major flares preceding the nebular phase (Extended Data Fig. 1).

The timescale of the optically thick (P Cyg) to optically thin (extended wings) transition is the diffusion timescale of an expanding shell with mass M_{shell} , and it is estimated²¹ to be $t_{\text{dif}} \approx 23 \text{ days} (1/R_{15}) (M_{\text{shell}}/M_{\odot})$, where M_{\odot} is the mass of the Sun and R_{15} is the radius of the spherical envelope in units of 10^{15} cm . For $t_{\text{dif}} = 0.002\text{--}0.1 \text{ days}$, which would be a conservative timescale relevant to the evolution of the BD value, we obtain $M_{\text{shell}} \approx (10^{-8}\text{--}10^{-5})M_{\odot}$. This is consistent with the black hole blowing away a substantial fraction of the matter stored in its large (mass of $\sim 10^{-5}M_{\odot}$) accretion disk²². On the other hand, this amount of mass is able to explain the increase in the equivalent hydrogen column density (of up to $N_{\text{H}} \approx 10^{24} \text{ cm}^{-2}$) observed during both the 1989 and the 2015 outbursts²² (Methods).

The active phase of the 2015 outburst of V404 Cyg is much shorter ($\sim 15 \text{ days}$) than typically observed in other luminous black holes (months to a year). This is followed by a sharp decay ($\sim 3 \text{ days}$), still during the radio-loud phase of the outburst, directly after the X-ray peak is reached. This behaviour is consistent with that observed in the 1989 outburst²². During these brief outbursts only about 0.1% (that is, $(0.3\text{--}1.1) \times 10^{-8}M_{\odot}$) of the material stored in the accretion disk is accreted by the black hole. This corresponds to the gas kept in the innermost $\sim (6\text{--}9) \times 10^5 \text{ km}$, which is unaffected by the long-lived outer disk wind. The amount of mass transferred from the donor star to the accretion disk during the preceding 26 years of quiescence is estimated to be $-\Delta M_2 \approx 3 \times 10^{-8}M_{\odot}$ (Methods). This amount is comparable to that accreted by the black hole and ejected by the wind. We also detect a prominent double-peaked H α line right after the end of the nebular phase, indicating the presence of a remnant accretion disk once the most active phase of the outburst was finished. Strong H α emission has in fact been observed throughout the inter-outburst interval, and it is

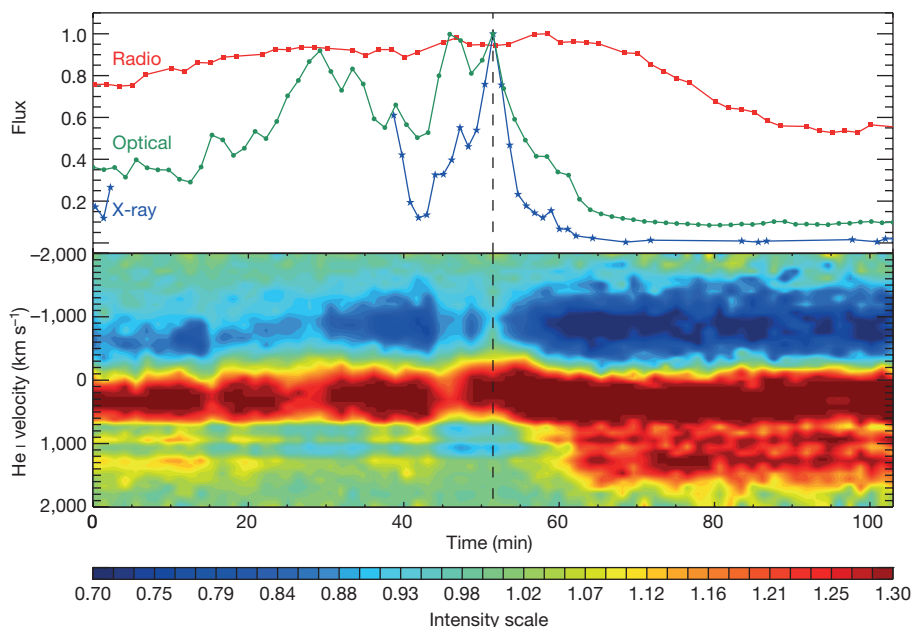


Figure 2 | Trailed spectrum corresponding to data from 19 June (day 2). The trail (bottom panel) covers 103 min with 75 spectra. Time corresponds to minutes from MJD 57,192.04. The normalized intensity scale is such that absorptions are represented in blue colours, while emissions are plotted in red colours. Simultaneous X-ray (Integral; blue stars), optical (green dots) and radio (red squares) normalized light curves are shown in the top

panel. Outflows are detected along the observation, but their properties change in response to flaring. The strongest features become evident directly after a sharp X-ray flare is seen (dashed line), as soon as the X-ray flux decreases and I_{ratio} reaches values as low as 0.5. During the flare (at ~ 0.08 times the flux peak observed later in the outburst), the P Cyg profile becomes weaker, as I_{ratio} increases to values larger than unity (up to ~ 2).

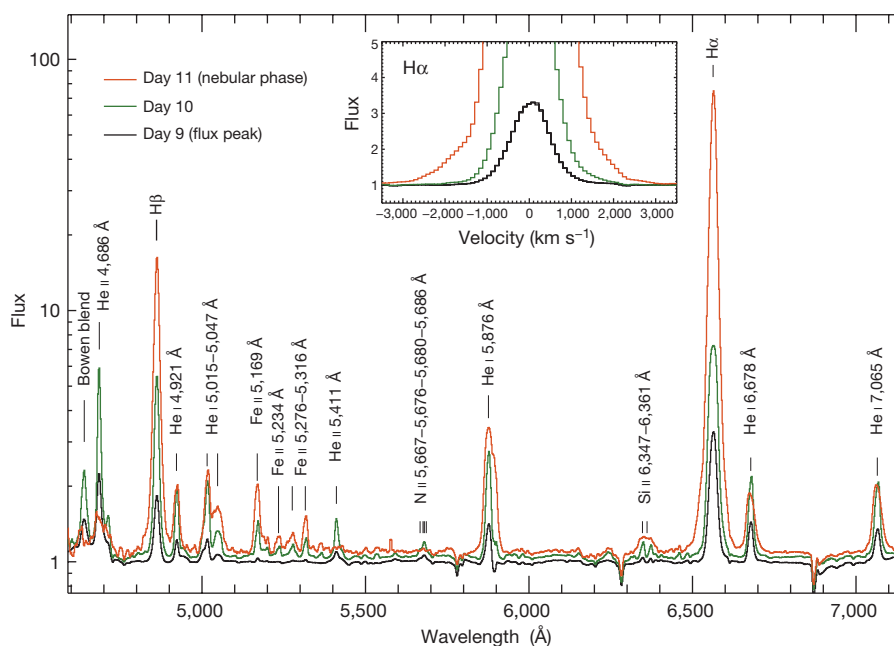


Figure 3 | Spectral evolution towards the nebular phase. Average, normalized spectra corresponding to days 9–11. A log scale has been used to represent the intense H α emission, which reached an equivalent width of 2,000 Å. An offset of 0.1 and 0.2 has been added to the day-10 and day-11 spectra, respectively. The optical flux drops by two orders of magnitude,

corresponding to the decay of the X-ray and radio outburst (Extended Data Fig. 1). He I and Balmer lines become intense and broad as other transitions become evident (Si II, Fe II). The inset shows the H α region, where broad wings reaching $\pm 3,000$ km s $^{-1}$ become apparent.

formed at $\sim 80\%$ of the outer disk radius²³ ($0.8R_{\text{out}} = 7 \times 10^6$ km). This (quiescent) period is also characterized by an X-ray luminosity about two orders of magnitude brighter than typically observed in quiescent black holes²⁴, implying ongoing accretion from the remnant accretion disk. Similarly, the much fainter secondary outburst detected in December 2015²⁵ also indicates the presence of an active disk only six months after the major outburst. This time lapse is consistent with the viscous timescale for refilling the inner disk (Methods).

In contrast with the sparse optical data obtained during the 1989 outburst²⁶, the intense observing campaign presented here allowed us to study in detail the evolution of the wind outflow and to detect the short-lived nebular phase. A nebular phase might also have occurred in the 1989 outburst—where intermittent P Cyg profiles were detected²⁶—but have been missed because of the scarce monitoring. On the other hand, the relative proximity of V404 Cyg enables both detailed spectroscopic observations at luminosities as low as $10^{-3}L_{\text{Edd}}$ and the detection of outflow features as weak as 2% of the continuum level during the brightest phases. In addition, the large accretion disk implied by the 6.5-day orbital period—the majority of black holes have orbital periods shorter than ~ 2 days (ref. 27)—provides the resource for the formation of outer disk outflows. Besides V404 Cyg, the behaviour of the other two systems with the longest orbital periods might also be influenced by the presence of mass outflows. V4641 Sagittarii, with $P_{\text{orb}} = 2.8$ days, has shown several brief outbursts characterized by strong radio emission. Extended H α wings ($\pm 2,500$ km s $^{-1}$) have been reported in a low-luminosity observation, possibly with a weak P Cyg profile in a Fe II emission line²⁸. Likewise, GRS 1915 + 105, the black hole with the longest orbital period, has been permanently in outburst for the past 23 years, alternating lower-luminosity plateau phases with short luminous episodes lasting only a few weeks⁴.

It is interesting to note that both V404 Cyg and GRS 1915 + 105 share distinctive variability patterns in their X-ray and optical emission, regardless of their differing luminosities⁷. These include short-term variations with large amplitudes, which, in addition to the neutral wind outflows reported here, seem to be a common feature of long-period black holes. This variability pattern has been proposed to result from insufficient mass flow reaching the inner parts of large disks, and

it might also affect the outburst evolution in addition or alternatively to the presence of the disk outflow presented here. The highly ionized wind of GRS 1915 + 105 during high-accretion-rate phases has been suggested to have a role in explaining the variability properties of the source, thereby linking (X-ray) outflows and oscillation patterns²⁹. Unfortunately, this system cannot be observed in the optical part of the spectrum owing to high interstellar extinction. Furthermore, it is not clear whether or not a similar coupling mechanism could be at work at the much lower luminosities associated with both the variability patterns⁷ and the neutral wind outflow observed in V404 Cyg.

The sustained disk wind that we have discovered in V404 Cyg could be a new fundamental driver in the accretion process of the largest, and hence most powerful, black-hole accretion disks. The outflow probably regulates the evolution of the outburst by depleting a sizeable fraction of the outer disk, thereby detaching the innermost regions, which are eventually accreted. This suggests behaviour analogous to that of the cold and massive outflows seen in active galactic nuclei, which shape their host galaxies at long distances from the central black hole³⁰.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 9 December 2015; accepted 22 February 2016.

Published online 9 May; corrected online 1 June 2016

(see full-text HTML version for details).

1. Fabian, A. C. Observational evidence of active galactic nuclei feedback. *Annu. Rev. Astron. Astrophys.* **50**, 455–489 (2012).
2. Fender, R. P. & Belloni, T. M. GRS 1915+105 and the disc-jet coupling in accreting black hole systems. *Annu. Rev. Astron. Astrophys.* **42**, 317–364 (2004).
3. Belloni, T. M., Motta, S. E. & Muñoz-Darias, T. Black hole transients. *Bull. Astron. Soc. India* **39**, 409–428 (2011).
4. Fender, R. P. & Muñoz-Darias, T. in *Astrophysical Black Holes. Lecture Notes in Physics*. Vol. 905, 65–100 (eds Haardt, F. et al.) (Springer, 2016).
5. Ponti, G. et al. Ubiquitous equatorial accretion disc winds in black hole soft states. *Mon. Not. R. Astron. Soc.* **422**, L11–L15 (2012).
6. Neilsen, J. & Lee, J. C. Accretion disk winds as the jet suppression mechanism in the microquasar GRS 1915+105. *Nature* **458**, 481–484 (2009).
7. Kimura, M. et al. Repetitive patterns in rapid optical variations in the nearby black-hole binary V404 Cygni. *Nature* **529**, 54–58 (2016).

8. Casares, J., Charles, P. A. & Naylor, T. A 6.5-day periodicity in the recurrent nova V404 Cygni implying the presence of a black hole. *Nature* **355**, 614–617 (1992).
9. Miller-Jones, J. C. A. The first accurate parallax distance to a black hole. *Astrophys. J.* **706**, L230–L234 (2009).
10. Barthelmy, S. D. *et al.* Swift trigger 643949 is V404 Cyg. *GRB Coord. Netw. Circ.* **17929** <http://gcn.gsfc.nasa.gov/gcn/gcn3/17929.gcn3> (2015).
11. Rodríguez, J. *et al.* Correlated optical, X-ray, and γ -ray flaring activity seen with INTEGRAL during the 2015 outburst of V404 Cygni. *Astron. Astrophys.* **581**, L9 (2015).
12. Martí, J., Luque-Escamilla, P. L. & García-Hernández, M. T. Multi-colour optical photometry of V404 Cygni in outburst. *Astron. Astrophys.* **586**, A58 (2016).
13. Muñoz-Darias, T. *et al.* Detection of transient optical P-Cygni profiles in V404 Cyg. *Astron. Telegr.* 7659 (2015).
14. Mauche, C. W. & Raymond, J. C. IUE observations of the dwarf nova HL Canis Majoris and the winds of cataclysmic variables. *Astrophys. J.* **323**, 690–713 (1987).
15. Castor, J. I. & Lamers, H. J. G. L. M. An atlas of theoretical P Cygni profiles. *Astrophys. J. Suppl. Ser.* **39**, 481–511 (1979).
16. Begelman, M. C., McKee, C. F. & Shields, G. A. Compton heated winds and coronae above accretion disks. I. Dynamics. *Astrophys. J.* **271**, 70–88 (1983).
17. King, A. L. *et al.* High-resolution Chandra HETG spectroscopy of V404 Cygni in Outburst. *Astrophys. J.* **813**, L37 (2015).
18. Drake, A. S. & Ulrich, R. K. The emission-line spectrum from a slab of hydrogen at moderate to high densities. *Astrophys. J. Suppl. Ser.* **42**, 351–383 (1980).
19. Iijima, T. & Esenoglu, H. H. Spectral evolution of Nova (V1494) Aql high velocity jets. *Astron. Astrophys.* **404**, 997–1009 (2003).
20. Thackeray, A. D. Spectra of the low-excitation nebulosities around AG Carinae and HD 138403. *Mon. Not. R. Astron. Soc.* **180**, 95–102 (1977).
21. Smith, N., Mauerhan, J. C. & Prieto, J. L. SN 2009ip and SN 2010mc: core-collapse Type II supernovae arising from blue supergiants. *Mon. Not. R. Astron. Soc.* **438**, 1191–1207 (2014).
22. Zycki, P. T., Done, C. & Smith, D. A. The 1989 May outburst of the soft X-ray transient GS 2023+338 (V404 Cyg). *Mon. Not. R. Astron. Soc.* **309**, 561–575 (1999).
23. Casares, J. A. FWHM- K_2 correlation in black hole transients. *Astrophys. J.* **808**, 80 (2015).
24. Armas Padilla, M. *et al.* Swift J1357.2–0933: the faintest black hole? *Mon. Not. R. Astron. Soc.* **444**, 902–905 (2014).
25. Beardmore, A. P., Page, K. L. & Kuulkers, E. Swift triggers on V404 Cyg. *Astron. Telegr.* 8455 (2015).
26. Casares, J., Charles, P. A., Jones, D. H. P., Rutten, R. G. M. & Callanan, P. J. Optical studies of V404 Cyg, the X-ray transient GS2023+338. I. The 1989 outburst and decline. *Mon. Not. R. Astron. Soc.* **250**, 712–725 (1991).
27. Corral-Santana, J. M. *et al.* BlackCAT: a catalogue of stellar-mass black holes in X-ray transients. *Astron. Astrophys.* **587**, A61 (2016).
28. Lindström, C. *et al.* New clues on outburst mechanism and improved spectroscopic elements of the black hole binary V4641 Sagittarii. *Mon. Not. R. Astron. Soc.* **363**, 882–890 (2005).
29. Nielsen, J., Remillard, R. A. & Lee, J. C. The physics of the Heartbeat State of GRS 1915+105. *Astrophys. J.* **737**, 69 (2011).
30. Feruglio, C. *et al.* Quasar feedback revealed by giant molecular outflows. *Astron. Astrophys.* **518**, L155 (2010).

Acknowledgements Nine of the spectra of 27 June were taken during the visit of King Felipe VI of Spain to the 10.4-m Gran Telescopio Canarias (GTC); we appreciate the support this visit provides to astrophysical research in Spain. This work is based on observations made with the GTC telescope, in the Spanish Observatorio del Roque de los Muchachos of the Instituto de Astrofísica de Canarias, during both Time Allocation Committee and Director's Discretionary observing time. We are thankful to the GTC team for the fast response and efficient work throughout the observing campaign. We acknowledge support by the Spanish Ministerio de Economía y competitividad under grants AYA2013-42627 and PSR2015-00397, the Leverhulme Trust Visiting Professorship Grant VP2-2015-046, the International Research Fellowship program of the Japan Society for the Promotion of Science (PE15024), the Bundesministerium für Wirtschaft und Technologie (BMW/DLR, FKZ 50 OR 1408) and the French Research National Agency's CHAOS project ANR-12-BS05-0009. The use of the MOLLY software developed by T. R. Marsh is gratefully acknowledged.

Author Contributions T.M.-D. performed the GTC data analysis and wrote the paper. J.C. contributed to the GTC data analysis and assisted in writing the paper. D.M.S. performed the GTC data reduction and contributed to the GTC data analysis. R.P.F. provided the radio data and contributed to the scientific discussion. M.A.P. performed X-ray analysis and contributed to the scientific discussion. M.L. provided day-12 GTC spectra and assisted in writing the paper. G.P. contributed to the scientific discussion. P.A.C. contributed to the scientific discussion and assisted in writing the paper. K.P.M. performed radio data analysis. J.R. provided part of the INTEGRAL data.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to T.M.-D. (teo.munoz-darias@iac.es).

METHODS

Optical spectroscopy. *Observations.* V404 Cyg was observed with the Optical System for Imaging and low-Intermediate-Resolution Integrated Spectroscopy (OSIRIS) located at the Nasmyth-B focus of the 10.4-m Gran Telescopio Canarias (GTC), La Palma (Spain). We used three different optical grisms; R1000B (1.12 Å per pixel), R2500R (1.04 Å per pixel) and R2500V (0.80 Å per pixel), which, combined with a 1.0'' slit, give $R = 611$, 1,485 and 1,509, respectively. These cover the spectral ranges: 3,630–7,500 Å, 5,575–7,685 Å and 4,500–6,000 Å. The complete observational set consists of 545 spectra obtained on 14 different nights within the period 17 June to 1 July (see Extended Data Table 1). The slit was rotated to $PA = 59.15^\circ$ to allow for simultaneous observations of a field star.

Data analysis. GTC spectra were bias- and flat-field-corrected using Image Reduction and Analysis Facility (IRAF) standard routines. The wavelength calibration was performed using Hg–Ar, Ne, and Xe lamps provided by the GTC team. Small velocity drifts ($<20 \text{ km s}^{-1}$) due to instrumental flexure were measured from the centroid of the $O \text{ I } \lambda = 5,577.340 \text{ Å}$ and $\lambda = 6,300.304 \text{ Å}$ lines and used to correct the individual spectra. We use routines within MOLLY (<http://deneb.astro.warwick.ac.uk/phsaap/software/molli/html/INDEX.html>) and IDL (Interactive Data Language) to perform further analysis of the spectra. Spectra were initially flux-calibrated relative to the comparison field star included in the slit. The latter was calibrated relative to the spectrophotometric standard star Wolf 1346, using a low-resolution spectrum taken with a 10-arcsec slit during a photometric night at air mass ~ 1.05 . Finally, we applied this flux calibration to the whole database. *Balmer decrement.* The BD was obtained by computing the flux ratio $H\alpha$ to $H\beta$ after subtracting the underlying continuum flux and correcting for reddening³¹ of $E_{B-V} = 1.3$. Results are comparable to those obtained from the ratio of the peak intensities of both lines, and results are consistent with case B recombination ($BD \approx 2.5\text{--}3$) with the exception of days 11 to 15, where values as high as 6 are observed. This epoch is the so-called nebular phase (see Extended Data Figs 1 and 5). We also note that values larger than ~ 3 are found corresponding to the strongest P Cyg profiles witnessed on days 1, 2 and 6.

Ionization state. Given the dramatic flux changes observed in the X-ray and optical bands, strong changes in the ionization state of the disk are expected as a result of differing irradiation levels. A good, widely used tracker of the variable irradiation of the outer disk is the $\text{He II } \lambda = 4,686 \text{ Å}$ to $H\beta$ line flux ratio (I_{ratio}) as it is reddening-independent. Nevertheless, as for the BD, we compute this ratio after subtracting the underlying continuum flux and correcting by $E_{B-V} = 1.3$. Not surprisingly, I_{ratio} is strongly correlated with the optical flux. Studies performed on accreting white dwarfs have shown some distinctive optical properties when I_{ratio} is larger than unity as a result of high ionization³². In our analysis we find that strong P Cyg profiles (blue absorption deeper than 5% of the continuum) are always associated with $I_{\text{ratio}} < 1$. The deepest profiles (30% below the continuum) are seen at $I_{\text{ratio}} = 0.5$.

$H\alpha$ profile. A visual inspection of the evolution of the $H\alpha$ line profile reveals a systematic asymmetry. We fitted the $H\alpha$ profile for each spectrum using a Gaussian model centred at the rest wavelength. We find that the line is redshifted during the first 11 days (top panel in Extended Data Fig. 4), reaching velocities above $\sim 100 \text{ km s}^{-1}$ during the first 8 days (that is, just before the brightest phase of the outburst), corresponding to the strongest P Cyg profiles. Similarly, the V/R ratio (defined as the ratio of the blue to red equivalent widths) confirms the line asymmetry observed up to day 11 (bottom panel in Extended Data Fig. 4). Error bars account for variability within every observing window. We interpret this as a result of continuous blue absorption (and extra red emission) present in the line profile, with the more extreme cases leading to P Cyg profiles. This is also sensitive to the presence of low-velocity outflows, which partially cover more central parts of the blue line profile.

X-ray observations. *Integral observatory.* V404 Cyg was extensively monitored by the INTEGRAL³³ satellite during the 15-day-long outburst starting on MJD 57,190 (17 June 2015)³⁴. The data were acquired during satellite revolutions 1,554 to 1,563. In Extended Data Fig. 1 we present a 25–200-keV light curve obtained with the Imager on Board the INTEGRAL Satellite (IBIS) and the upper-layer detector INTEGRAL Soft Gamma-Ray Imager (ISGR) in 64-s time bins³⁵.

The raw data were reduced in a very standard way, using the Off-line Scientific Analysis software (OSA) version 10.1 (http://www.isdc.unige.ch/integral/download/osa/doc/10.0/osa_inst_guide.pdf) similar to that described in the case of V404 Cyg¹¹ and Cyg X-1³⁶ (same region of the sky), respectively. IBIS is a coded mask telescope and the data reduction process is iterative: each active source in a given field projects its own shadow onto the detector, and hence contributes to the overall background of the other sources. Hence, to extract scientific products of one specific object one must consider all other active/bright sources within the field.

Our reduction procedure started with the production of sky images and mosaics (obtained from combining data acquired during the same satellite revolution)

in the user's pre-defined energy ranges (here we used 20–40 keV, 40–80 keV, 80–150 keV and 150–300 keV) to identify the most active sources over each revolution. In this region of the sky (Cygnus) two persistent sources are very bright and active in the hard X-ray/soft γ -ray domain ($\sim 0.2\text{--}1 \text{ MeV}$), Cygnus X-1 and Cygnus X-3³⁶, in addition to V404 Cyg. Occasionally, some other objects may show up (for example, EXO 2030+375), and are thus included in the reduction process of the revolution concerned.

Light curves were then extracted in time steps of $\sim 64 \text{ s}$ over the 25–60-keV and 60–200-keV energy ranges. These energy bands are the same as those selected by the INTEGRAL Science Data Centre quick look analysis facility (<http://www.isdc.unige.ch/integral/analysis#QLAsources>), providing an independent check and also avoiding any potential saturation effects in the 20–25-keV energy range usually considered.

Swift observatory. Following the initial alert¹⁰, V404 Cyg was monitored with the Swift satellite throughout its outburst until it returned to quiescence. We have analysed a total of 43 observations acquired with the X-ray Telescope³⁷ (XRT) taken from 17 days starting on MJD 57,188 (15 June 2015).

A total of 35 observations were performed in the windowed timing mode, while 8 were taken in the photon counting mode. Observations were processed using the HEASOFT v.6.17 software, in particular the XRTPIPELINE task. For each observation the 0.5–10-keV spectrum, light curve and image were obtained using XSELECT. We used a circular region of 40-arcsec radius centred at the source position (the inner $\sim 9\text{--}11$ arcsec were excluded for those observations affected by pile-up). A region of similar size and shape, positioned on an empty sky region, was used for the background. We created exposure maps and ancillary response files following the standard Swift analysis threads (<http://www.swift.ac.uk/analysis/xrt/>), and we acquired the last version of the response matrix files from the High Energy Astrophysics Science Archive Research Center (HEASARC) calibration database (CALDB).

Radio observations. V404 Cyg was observed extensively throughout its 2015 outburst between 13 GHz and 18 GHz by the AMI-LA radio telescope (Cambridge, UK), operating as part of the University of Oxford's 4 PI SKY (<http://www.4pisky.org>) transients programme. The first data were obtained on MJD 57,188.896 (15 June 2015) in robotic response to the Swift trigger 643949. The observations took place within two hours of the Swift trigger, revealing a bright ($>100 \text{ mJy}$) and fading radio flare³⁸. Subsequently, we continued to monitor the source for up to 10 h every day for the entire period covered by this report.

Quick-look images of the AMI-LA observations were obtained with the fully-automated calibration and imaging pipeline, AMISURVEY³⁹. After the outburst, a more careful calibration and radio-frequency interference excision of the raw data was done using AMI-REDUCE⁴⁰. The calibrated data was then imported to the CASA package^{41,42}. Light curves were extracted in time steps of $\sim 1 \text{ s}$ in six channels across the 5-GHz bandwidth via vector-averaging of the UV-plane data.

Over the period of 15 days of maximum activity of V404 Cyg, flares with peak flux density of up to 3 Jy at 16 GHz are seen. The rise of the flares is generally optically thick, while the decay is optically thin, consistent with adiabatically expanding blobs of plasma (constituting the jet).

Fundamental parameters of V404 Cyg. V404 Cyg is a dynamically confirmed black hole X-ray binary with an orbital period of $P_{\text{orb}} = 6.47$ days (ref. 8). The black-hole mass is in the range $M_{\text{BH}} = (8\text{--}12)M_{\odot}$, the error budget being dominated by the uncertainty in the orbital inclination⁴³. We use a black-hole mass of $M_{\text{BH}} = 10M_{\odot}$ and a mass ratio of $q = M_{\text{BH}}/M_2 = 0.067$ in every calculation presented in this paper⁴³, where M_2 corresponds to the donor-star mass in units of M_{\odot} . This results in an orbital separation of $2.2 \times 10^7 \text{ km}$. The outer accretion disk radius R_{out} (in units of kilometres) can be expressed as⁴⁴:

$$R_{\text{out}} \cong 1.2 \times 10^6 M_{\text{BH}}^{\frac{1}{3}} P_{\text{orb}}^{\frac{2}{3}}$$

where M_{BH} is expressed in units of M_{\odot} and P_{orb} in days. We obtain $R_{\text{out}} = 9 \times 10^6 \text{ km}$. We note that the vast majority of black holes have orbital periods shorter than ~ 2 days, which results in $R_{\text{out}} = 4.1 \times 10^6 \text{ km}$ if we use the same values as for V404 Cyg for M_{BH} and q .

P Cyg profiles. We discovered strong P Cyg profiles in both the hydrogen (Balmer) and helium (He I) emission lines. They result from resonant scattering in an expanding outflow, and are well reproduced by models using a variety of velocity laws and a spherical geometry^{14,15}.

Profile fitting. To constrain the velocities associated with the He I ($\lambda = 5,876 \text{ Å}$) P Cyg profile, we fitted every individual spectrum as follows:

(1) We fitted a Gaussian to the central disk emission after masking both the blue P Cyg absorption and the red high-velocity emission bump. This fitted model was subtracted from the data.

(2) We subsequently fitted a two-Gaussian model to the residuals, one in absorption to the P Cyg and another in emission to the red bump. To avoid degeneracy in the fit, both Gaussians were offset by the same velocity (sign understood) and set to have the same width. The intensities were left as free parameters.

Fits provide a good description of the data for relatively deep profiles, as indicated by our visual inspections. We take the wind mean velocity to be the offset velocity, while V_T (the wind terminal velocity) is determined by adding to the wind mean velocity the half-width at 1/10th of the intensity. For days 1, 2 and 6, where strong profiles are present throughout the observations, we were able to track their evolution with time. On day 2 (Fig. 2) outflows are detected during the 2 h observation, but their properties—profile strength and velocity—change in response to flaring. Gaussian fits show a constant V_T throughout the observation, while both the amplitude and mean velocity vary following changes in the optical flux and ionization state. A similar flare–outflow correlation is witnessed on day 6, where we measure $V_T = 3,000 \text{ km s}^{-1}$.

Possible physical interpretations. A variety of wind-launching mechanisms have been proposed in the literature. Here, we briefly discuss the three more widely used. (1) Radiation pressure. The wind results from Thompson scattering when the radiation field approaches L_{Edd} . Given the low luminosity at which the most conspicuous P Cyg profiles are detected (0.1%–1% of L_{Edd}), and the similar V_T values observed during the outburst, radiation pressure is probably not responsible for the observed phenomenology. However, the system might have reached L_{Edd} during the brightest phases of the outburst preceding the so-called nebular phase, and this mechanism could have contributed to the observed optically thin shell. On the other hand, we note that the velocity observed during this phase is similar to that measured in the P Cyg profiles, suggesting a common origin.

(2) Line-driven winds. Such winds are expected to be inefficient in low-mass X-ray binaries since X-ray emission from the disk would over-ionize the wind⁴⁵.

(3) Thermal wind scenario¹⁶. In this scenario, atoms reach a thermal velocity larger than the escape velocity and a wind is formed. Therefore, the launching radius R_l is approximately that for which an associated Keplerian velocity equals V_T . For $V_T = 3,000$ – $1,500 \text{ km s}^{-1}$ we obtain $R_l = (1.5\text{--}6) \times 10^5 \text{ km}$, respectively.

Using a standard accretion disk model⁴⁶ we can estimate the surface temperature of the disk at a given radius R by:

$$T_D(R) = \left[\frac{3GM_{\text{BH}}M_{\text{acc}}}{8\pi R^3\sigma} \left(1 - \sqrt{\frac{R_0}{R}} \right) \right]^{1/4}$$

where G is the gravitational constant, σ is the Stefan–Boltzmann constant, and R_0 is the disk inner radius $R_0 = \frac{6GM_{\text{BH}}}{c^2}$.

The accretion rate is obtained using:

$$\dot{M} \approx 2.0 \times 10^{39} \frac{L_{\text{out}}}{\eta c^2}$$

where L_{out} is the Eddington scaled luminosity at the time of the outflow, c is the speed of light and $\eta = 0.1$ is the accretion efficiency. Using $L_{\text{out}} = (0.001\text{--}0.01)L_{\text{Edd}}$, we obtain T_D in the range 5,000–20,000 K, which is consistent with having both neutral hydrogen and helium.

Mass outflow rate. A crude estimate of the mass loss can be obtained by direct comparison of the strongest profiles (day 2) with the classical atlas of theoretical P Cyg profiles¹⁵. Following studies of cataclysmic variables⁴⁷ we calculated the rate of mass outflow (\dot{M}_{out}) (in units of solar masses per year) as follows:

$$\dot{M}_{\text{out}} \approx 1.1 \times 10^{-18} \times \frac{\tau R V_T^2}{f_i A \lambda_0 g} C$$

where τ is the average opacity, λ_0 is the rest wavelength, f_i is the ionization fraction, A is the helium abundance, R is the radius of the emitting region (in solar radii) and g is the oscillator strength. C is an integral depending on the shape of the P Cyg profile, which takes typical values in the range 0.2–0.5. By visual inspection we set $\tau \approx 2$ by comparing our profiles with those of the atlas.

Using $V_T = 2,000 \text{ km s}^{-1}$, $g = 0.61$, $C = 0.2$ and $A = 0.08$ we obtain $\dot{M}_{\text{out}} > 1 \times 10^{-14} M_{\odot} \text{ yr}^{-1}$. Note that this is a lower limit because (1) we have assumed that $f_i = 0.5$, which might be much lower depending on, for example, the exact value of the wind-launching radius (although the low He II ($\lambda = 4,686 \text{ \AA}$)/H3 intensity ratio $I_{\text{ratio}} = 0.5$ advocates for a substantial fraction of neutral helium), and (2) we used $R = 6 \times 10^5 \text{ km}$ for the proposed launching radius, which could be much larger if, for example, the shell remains optically thick $\sim 300 \text{ s}$ (0.003 days) after the outflow is launched. Indeed, if the shell is optically thick for at least 0.01 days (see below), we obtain $\dot{M}_{\text{out}} > 10^{-13} M_{\odot} \text{ yr}^{-1}$ (for $C = 0.5$).

On the other hand, the contemporaneous Eddington-scaled accretion rate for $L_{\text{out}} = 0.001 L_{\text{Edd}}$ would be $\dot{M}_{\text{acc}} \approx 10^{-10} M_{\odot} \text{ yr}^{-1}$ and then $\frac{\dot{M}_{\text{out}}}{\dot{M}_{\text{acc}}} > 10^{-3}$.

Nebular phase. From days 9 to 11 we observe major changes in the spectrum, as the X-ray, radio and optical fluxes drop by 2–3 orders of magnitude from the outburst peak: (1) A weak P Cyg profile is still present on days 9 and 10, which means that the ejecta are optically thick; (2) higher-excitation emission lines become strong on day 10 and Balmer line equivalent widths start to increase; (3) on day 11, emission lines become unprecedentedly broad and intense, showing zero-intensity breadths of $\sim 6,000 \text{ km s}^{-1}$ and equivalent widths of $\sim 2,000 \text{ \AA}$ (H α ; Fig. 3, Extended Data Fig. 1). This results from material expanding at the outflow velocity ($\pm 3,000 \text{ km s}^{-1}$) and becoming optically thin, as typically observed in expanding nova shells. We note that H α saturated the detector on day 10, so its intensity has to be taken as a lower limit. Similar spectra to that presented in Fig. 3 (day 11) were observed during days 12, 13 and 14, although line intensities progressively decay. Day 15 data show features typical of quiescent black hole transients, including a double peaked H α emission line.

Diffusion timescale and ejected mass. The timescale of this transition is the diffusion timescale of an expanding shell with mass M_{shell} , and it is estimated^{21,48} to be $t_{\text{dif}} \approx 23 \text{ days} (1/R_{15})(M_{\text{shell}}/M_{\odot})$, where R_{15} is the radius of the spherical envelope in units of 10^{15} cm . The transition from optically thick to optically thin ejecta occurs between days 10 and 11, as we observe a P Cyg profile and $\text{BD} \approx 2.5$ on day 10 and broad wings and $\text{BD} \approx 5$ on day 11 (Extended Data Fig. 5). This means that the outflow becomes optically thin in $t_{\text{dif}} < 1 \text{ day}$ if ejection of matter continues up to day 10. On the other hand, on day 11 we have two separate groups of observations with BD increasing across ~ 0.01 -day timescales, which suggests that this timescale could be relevant in the expansion. Extrapolating from this variation, we predict a maximum $t_{\text{dif}} \approx 0.1 \text{ day}$. On the other hand, we do not observe substantial changes from spectrum to spectrum ($t_{\text{dif}} > 0.002$). Assuming a $3 \times 10^5 \text{ km}$ launching radius and material travelling at $3,000 \text{ km s}^{-1}$ we obtain $R_{15} = 9 \times 10^{-5}$ to 3×10^{-3} . This yields $M_{\text{shell}} = (10^{-8}\text{--}10^{-5}) M_{\odot}$ for $t_{\text{dif}} \approx 0.002\text{--}0.1$, respectively. This order-of-magnitude calculation is consistent with the amount of matter expected to be stored in a large accretion disk such as that of V404 Cyg ($\sim 10^{-5} M_{\odot}$)²². Similarly, it also explains nicely why the optically thin nebulae is detected right after the end of the outburst ($< 1 \text{ day}$), a much shorter timescale than typically observed in supernovae (tens to hundreds of days) where $> (1\text{--}10) M_{\odot}$ are expelled²¹. The above estimates are quite approximate (at least a factor of ~ 2) and they assume a spherical geometry. Nevertheless, our results are consistent with a substantial fraction of the disk being ejected during the outburst. On the other hand, an increase in the equivalent hydrogen column density (N_{H} of up to a few times 10^{23}) has been reported for both the 1989²² and 2015⁴⁹ outbursts. Assuming a spherical geometry for the wind and constant density across the outflow, we predict $N_{\text{H}} \approx 10^{21}\text{--}10^{24} \text{ cm}^{-2}$ if $M_{\text{shell}} \approx (10^{-8}\text{--}10^{-5}) M_{\odot}$ were expelled.

Mass transferred by the donor during the inter-outburst period. Given the long orbital period of V404 Cyg, the mass transfer rate from the donor \dot{M}_2 (in units of solar masses per year) can be estimated using the following expression⁴⁴:

$$-\dot{M}_2 \approx 4.0 \times 10^{-10} P_d^{0.93} \dot{M}_2^{1.47}$$

where P_d is the orbital period in days. This results in $-\dot{M}_2 \approx 1.3 \times 10^{-9} M_{\odot} \text{ yr}^{-1}$, which translates into $-\Delta M_2 \approx 3 \times 10^{-8} M_{\odot}$ across the 26-yr-long inter-outburst period.

Accreted mass estimate. The total mass accreted during the 2015 outburst based on the observed X-ray luminosity can be estimated using the following expression:

$$\Delta M_X \approx \frac{\int L_X}{\eta c^2}$$

where $\int L_X$ is the integrated X-ray luminosity throughout the outburst. This was estimated by converting the observed Integral count rate to flux in the 10 keV to 1 MeV band (<https://heasarc.gsfc.nasa.gov/cgi-bin/Tools/w3pimms/w3pimms.pl>). We assumed¹¹ a power-law spectral model with a photon index in the range $\Gamma = 1\text{--}2$ and $N_{\text{H}} \approx 0.7 \times 10^{22}$, even though the result does not depend on the N_{H} value. Using $\eta = 0.1$ we obtain $\Delta M_X \approx (0.6\text{--}2.3) \times 10^{25} \text{ g}$, that is, $(0.3\text{--}1.1) \times 10^{-8} M_{\odot}$. The mass that is implied by the soft X-ray luminosity (0.5–10 keV), and that is therefore sensitive to (variable) absorption effects, is estimated from both Swift and Integral (extrapolation) to be only a few times 10^{23} g . Our results are compatible with the value of $\Delta M_X \approx 3 \times 10^{25} \text{ g}$ inferred during the 1989 outburst²², showing that only $\sim (0.5\text{--}1) \times 10^{-3}$ of the total mass stored in the disk ($R_{\text{out}} = 9 \times 10^6 \text{ km}$) was accreted. The disk mass^{22,50} varies as $\sim R_{\text{out}}^3$, which in turn implies that the accreted mass corresponds to that within $R_{\text{acc}} \approx (6\text{--}9) \times 10^5 \text{ km}$.

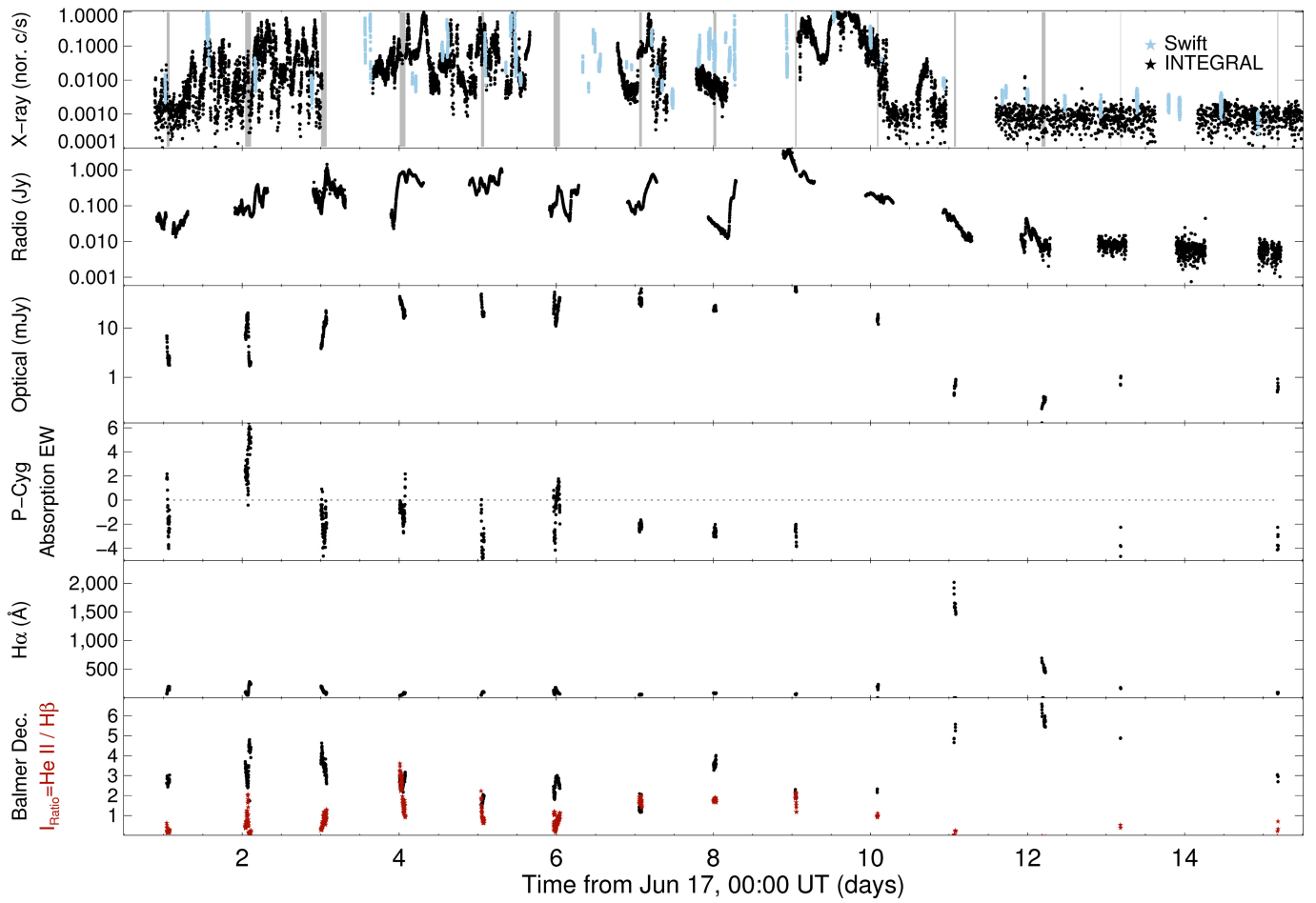
Renewed activity of V404 Cyg^{12,51}, at a much fainter level, was detected during the period 23 December 2015 to 5 January 2016, that is, only about six months

after the main 2015 outburst. This timescale is consistent with the viscous time for refilling R_{acc} as compared with accreting white dwarfs. This can be expressed as:

$$t_V \approx N_{\text{WD}} \times 2.87 \times 10^7 R_{10}^{0.61} M_{\text{BH}}^{0.46}$$

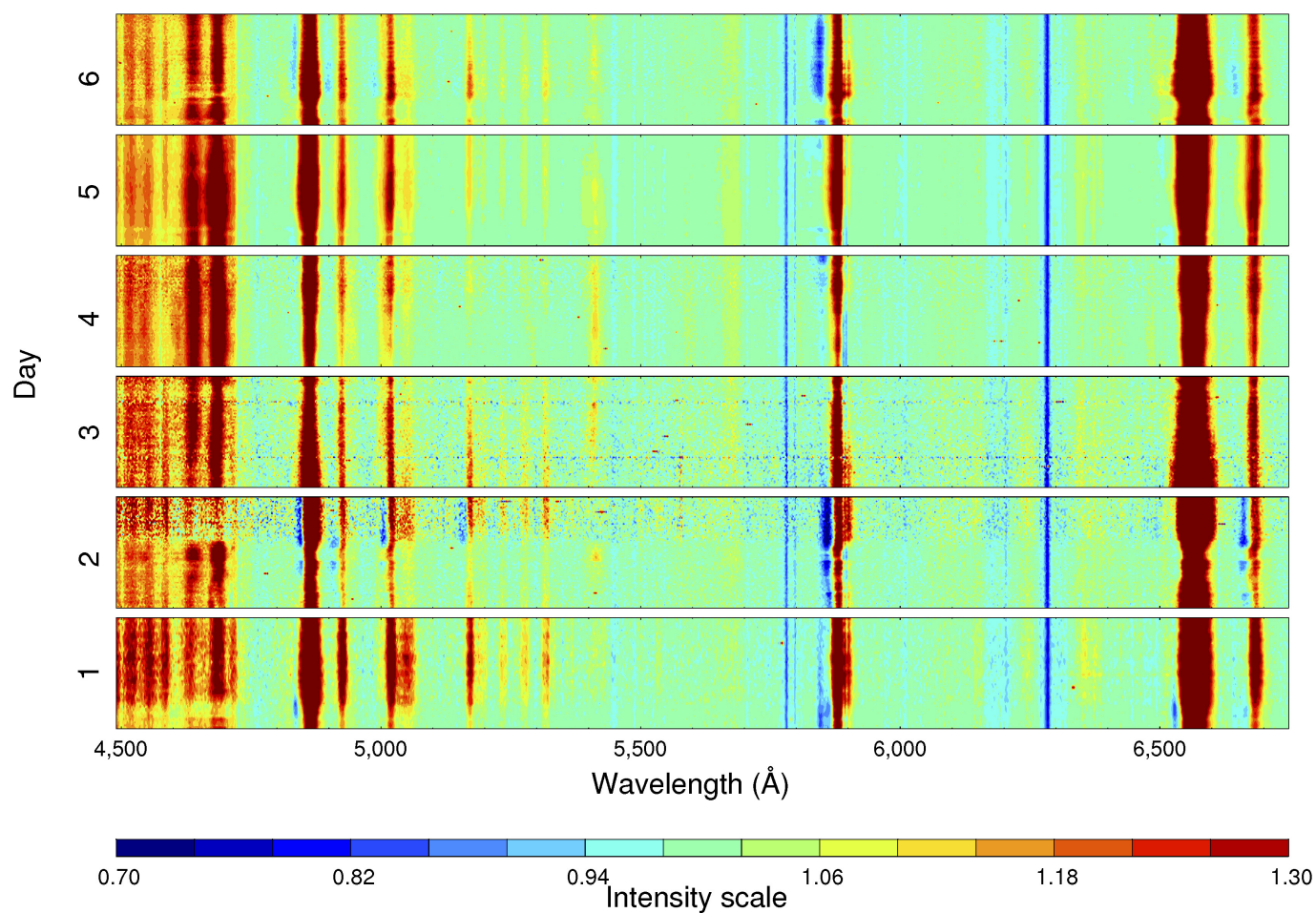
where R_{10} is the radius in units of 10^{10} cm and N_{WD} was (roughly) calibrated to be ~ 0.05 using observations of accreting white dwarfs⁵². Using $R_{\text{acc}} \approx (6-9) \times 10^5$ km, we obtain $t_V \approx 140-180$ days, which is compatible with the time lapse between the two outbursts.

31. Chen, W., Schrader, C. R. & Livio, M. The properties of X-ray and optical light curves of X-ray novae. *Astrophys. J.* **491**, 312–338 (1997).
32. Groot, P. J., Rutten, R. G. M. & van Paradijs, J. SW Sextantis in an excited, low state. *Astron. Astrophys.* **368**, 183–196 (2001).
33. Winkler, C. *et al.* The INTEGRAL mission. *Astron. Astrophys.* **411**, L1–L6 (2003).
34. Kuulkers, E. INTEGRAL observations of V404 Cyg (GS 2023+338): public data products *Astron. Telegr.* 7758 (2015).
35. Lebrun, F. *et al.* ISGRI: the INTEGRAL Soft Gamma-Ray Imager. *Astron. Astrophys.* **411**, L141–L148 (2003).
36. Rodriguez, J. *et al.* Spectral state dependence of the 0.4–2 MeV polarized emission in Cygnus X-1 seen with INTEGRAL/IBIS, and links with the AMI radio data. *Astrophys. J.* **807**, 17 (2015).
37. Burrows, D. N. *et al.* The Swift X-ray telescope. *Space Sci. Rev.* **120**, 165–195 (2005).
38. Mooley, K. P. *et al.* Bright radio flaring from V404 Cyg detected by AMI-LA. *Astron. Telegr.* 7658 (2015).
39. Staley, T. D. & Anderson, G. E. Chimenea and other tools: automated imaging of multi-epoch radio-synthesis data with CASA. *Astron. Comput.* **13**, 38–49 (2015).
40. Davies, M. L. *et al.* Follow-up observations at 16 and 33GHz of extragalactic sources from WMAP 3-yr data: I—spectral properties. *Mon. Not. R. Astron. Soc.* **400**, 984–994 (2009).
41. McMullin, J. P., Waters, B., Schiebel, D., Young, W. & Golap, K. CASA architecture and applications. *Astron. Soc. Pacif. Conf. Ser.* **766**, 127 (2007).
42. CASA Consortium. CASA: Common Astronomy Software Applications. *Astrophys. Source Code Lib.* ascl:1107.013, <https://casa.nrao.edu> (2011).
43. Casares, J. & Jonker, P. G. Mass measurements of stellar and intermediate-mass black holes. *Space Sci. Rev.* **183**, 223–252 (2014).
44. King, A. R., Kolb, U. & Burderi, L. Black hole binaries and X-ray transients. *Astrophys. J.* **464**, L127–L130 (1996).
45. Proga, D. & Kallman, T. R. On the role of the ultraviolet and X-ray radiation in driving a disk wind in X-ray binaries. *Astrophys. J.* **565**, 455–470 (2002).
46. Shakura, N. I. & Sunyaev, R. A. Black holes in binary systems. Observational appearance. *Astron. Astrophys.* **24**, 337–355 (1973).
47. Krautter, J. *et al.* IUE spectroscopy of cataclysmic variables. *Astron. Astrophys.* **102**, 337–346 (1981).
48. Smith, N. & Arnett, W. D. Preparing for an explosion: hydrodynamic instabilities and turbulence in presupernovae. *Astrophys. J.* **785**, 82 (2014).
49. Walton, D. *et al.* NuSTAR observation of V404 Cyg during/after decline. *Astron. Telegr.* 7752 (2015).
50. King, A. R. & Ritter, H. The light curves of soft X-ray transients. *Mon. Not. R. Astron. Soc.* **293**, L42–L48 (1998).
51. Motta, S. E. *et al.* INTEGRAL and Swift observations of V404 Cyg: going back to quiescence? *Astron. Telegr.* 8510 (2016).
52. Cannizzo, J. K., Shafter, A. W. & Wheeler, J. C. On the outburst recurrence time for the accretion disk limit cycle mechanism in dwarf novae. *Astrophys. J.* **333**, 227–235 (1988).

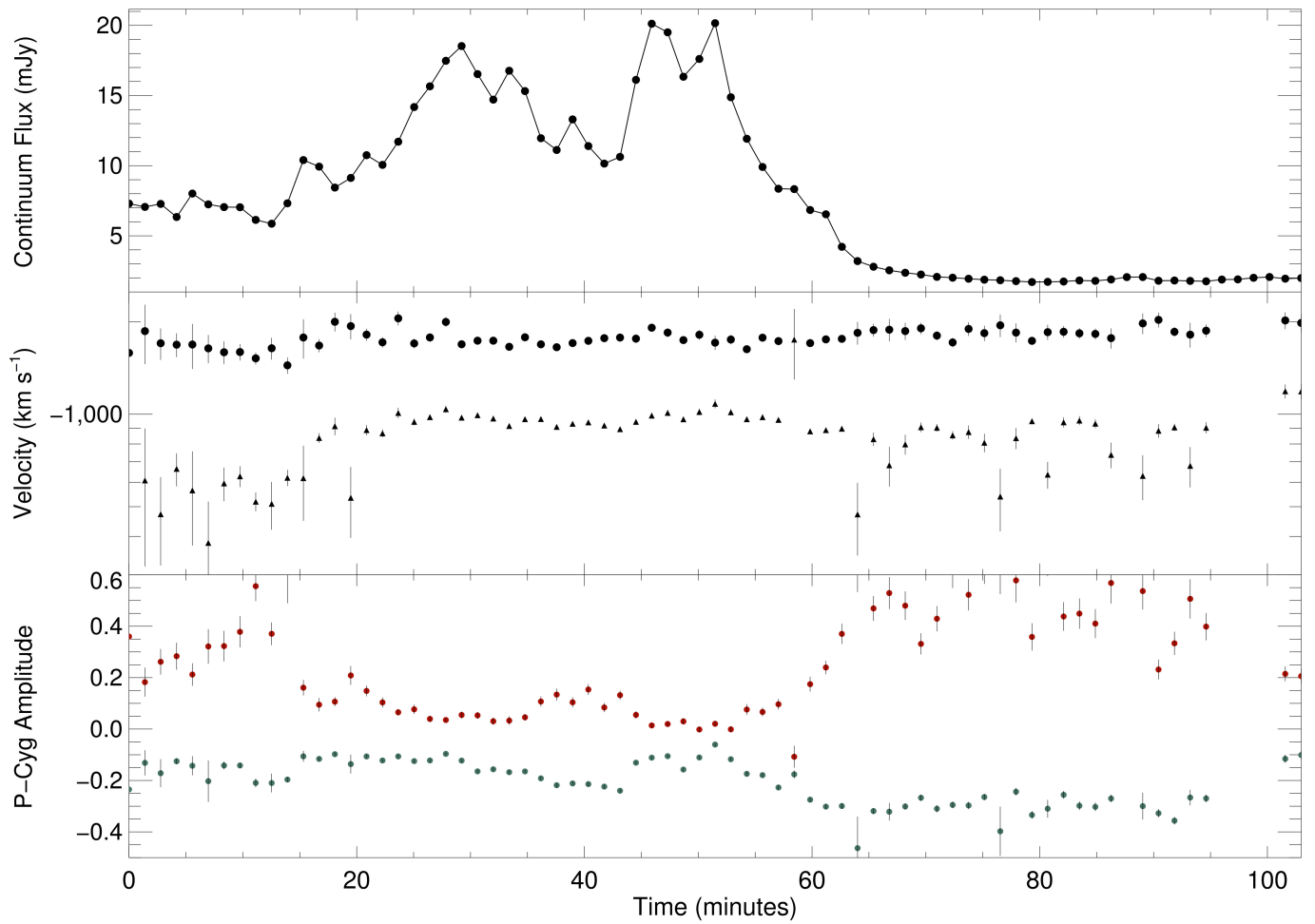


Extended Data Figure 1 | Evolution of the main parameters during the outburst. Zero time is set to 17 June 00:00 UT. From top to bottom we show hard (25–200 keV) and soft (0.5–10 keV) normalized X-ray count rates, radio flux (~ 16 GHz), optical continuum flux, He I $\lambda = 5,876$ Å equivalent width (EW; positive for absorption) in the range $-3,000$ km s $^{-1}$

to 0 km s $^{-1}$ and H α equivalent width. The Balmer decrement (black) and I_{ratio} (red) are shown in the bottom panel. In the top panel, X-rays have been normalized to their respective peak at $\sim L_{\text{Edd}}$ and the time intervals corresponding to the GTC observations have been greyed out for clarity.

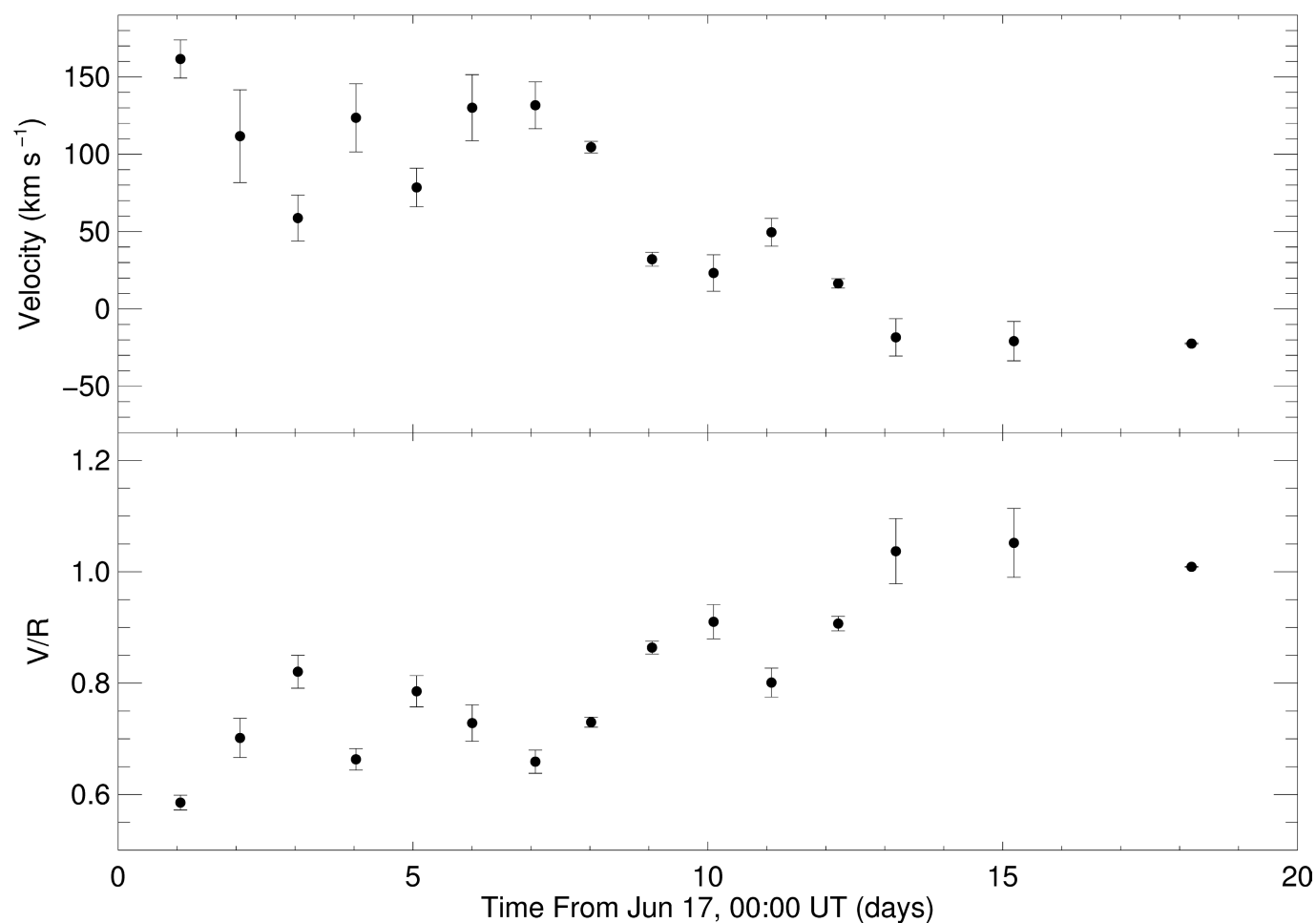


Extended Data Figure 2 | Trail spectrum showing GTC spectra taken during days 1–6. P Cyg profiles are apparent in seven transitions of neutral hydrogen ($H\alpha$ and $H\beta$) and helium. The strongest are observed in days 1, 2 and 6, being more prominent in the $He\ I\ \lambda = 5,876\ \text{\AA}$ transition. Similar profiles are seen in another five transitions at shorter wavelengths (not shown).



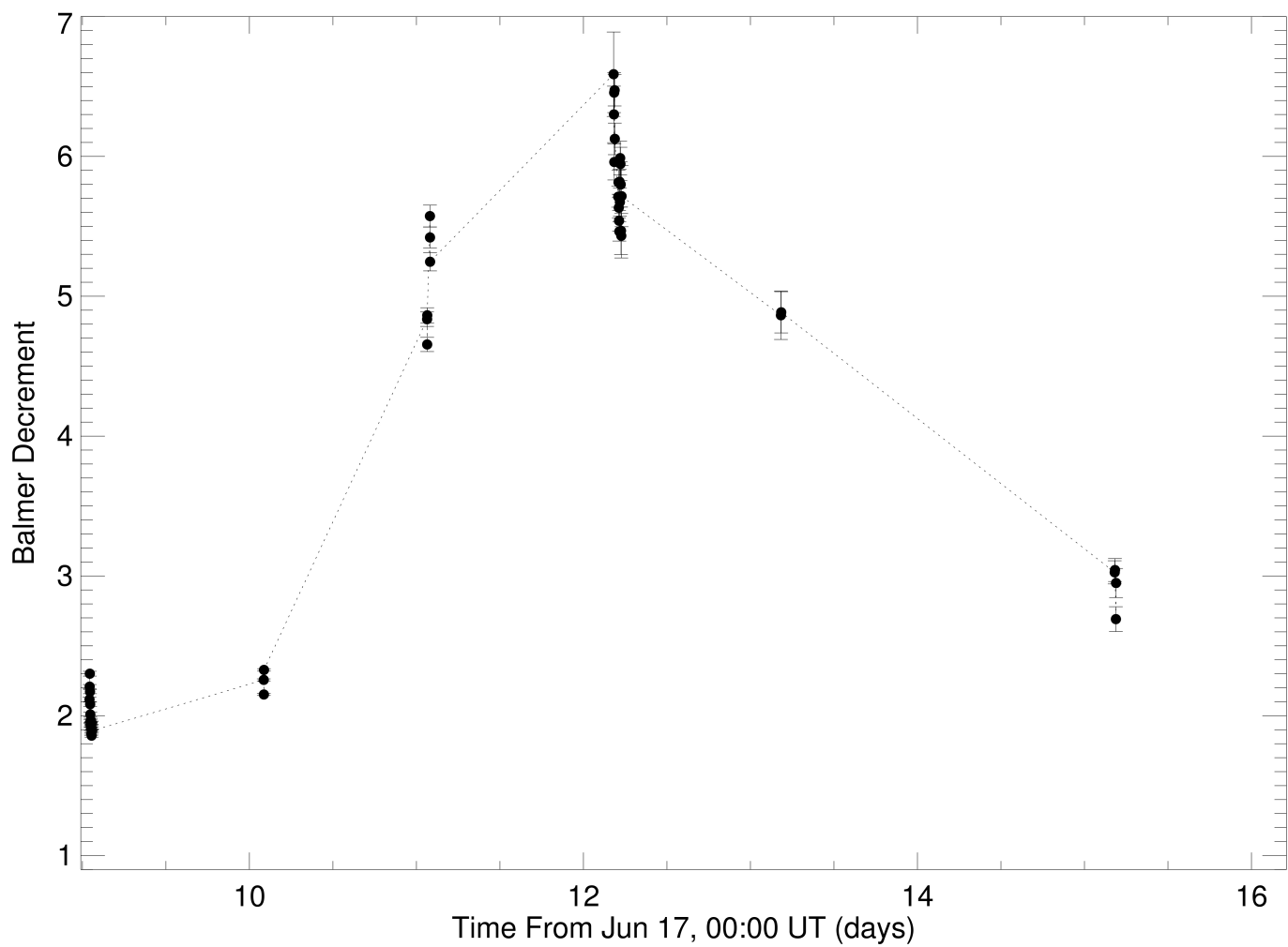
Extended Data Figure 3 | Gaussian fits to the P Cyg profiles. From top to bottom we show the optical flux, terminal (dots) and mean (triangles) velocities, and emission (red dots) and absorption amplitudes (blue dots) of the P Cyg profiles. The two bottom panels result from a Gaussian fitting after subtracting the disk component from the emission (Methods). Terminal velocities in the range $V_T = 1,500\text{--}2,000\text{ km s}^{-1}$ are observed

(see Fig. 2). This method yields $V_T = 3,000\text{ km s}^{-1}$ for day 6 (see Fig. 1). The amplitude of the profile is correlated with the optical flux. The similar (if not higher) amplitude of the emission component implies that the wind has a large covering factor. Error bars indicate the standard error of the mean.



Extended Data Figure 4 | Evolution of the H α emission line. The top panel shows the evolution of the centroid of the H α line. Positive velocity values are due to line asymmetries by blue absorption and red emission. The bottom panel shows the V/R parameter (the emission line

is symmetric if $V/R = 1$) showing the same phenomena. This strongly suggests the presence of continuous outflows from the outer disk along the whole outburst. Error bars indicate the standard deviation of measurements within each observing window.



Extended Data Figure 5 | BD evolution through the nebular phase. From day 10, the BD is observed to increase sharply, reaching ~ 5 on day 11 and ~ 6 on day 12 (see Methods). Dotted lines join observations consecutive in time. Error bars indicate the standard error of the mean.

Extended Data Table 1 | Log of the GTC observations

Night	Grism	T_{EXP} [s]	TR [s]	N_{SPEC}
17-06-2015	R1000B	60	84	36
18-06-2015	R1000B	60	84	75
19-06-2015	R1000B	60	84	75
20-06-2015	R1000B	60	84	75
21-06-2015	R1000B	60	84	40
22-06-2015	R1000B	60	84	85
23-06-2015	R1000B	60	84	36
24-06-2015	R1000B	60	84	36
25-06-2015	R1000B	60	84	17
26-06-2015	R1000B	40	64	3
	R2500R	70	94	6
	R2500V	70	94	6
27-06-2015	R1000B	20	44	3
	R2500V	70	94	3
	R2500R	35	59	3
	R1000B	20	44	3
	R2500V	70	94	3
	R2500R	35	59	3
28-06-2015	R1000B	20	-	1
	R1000B	60	84	5
	R2500V	360	384	3
	R2500R	120	144	3
	R1000B	120	144	13
29-06-2015	R1000B	120	144	2
	R2500R	120	144	2
01-07-2015	R1000B	120	144	2
	R2500R	120	144	2
	R1000B	60	84	2
	R2500R	60	84	2

T_{EXP} is the exposure time per spectrum in seconds, TR is the actual time resolution and N_{SPEC} is the number of spectra taken on a given day and with a given configuration.

Vigorous convection as the explanation for Pluto's polygonal terrain

A. J. Trowbridge¹, H. J. Melosh^{1,2}, J. K. Steckloff² & A. M. Freed¹

Pluto's surface is surprisingly young and geologically active¹. One of its youngest terrains is the near-equatorial region informally named Sputnik Planum, which is a topographic basin filled by nitrogen (N₂) ice mixed with minor amounts of CH₄ and CO ices¹. Nearly the entire surface of the region is divided into irregular polygons about 20–30 kilometres in diameter, whose centres rise tens of metres above their sides. The edges of this region exhibit bulk flow features without polygons¹. Both thermal contraction and convection have been proposed to explain this terrain¹, but polygons formed from thermal contraction (analogous to ice-wedges or mud-crack networks)^{2,3} of N₂ are inconsistent with the observations on Pluto of non-brittle deformation within the N₂-ice sheet. Here we report a parameterized convection model to compute the Rayleigh number of the N₂ ice and show that it is vigorously convecting, making Rayleigh–Bénard convection the most likely explanation for these polygons. The diameter of

Sputnik Planum's polygons and the dimensions of the 'floating mountains' (the hills of water ice along the edges of the polygons) suggest that its N₂ ice is about ten kilometres thick. The estimated convection velocity of 1.5 centimetres a year indicates a surface age of only around a million years.

Previous work first proposed that convection or thermal contraction could have formed the polygons on Sputnik Planum¹ (see Fig. 1). However, we find contraction unlikely: studies of Arctic ice-wedges show that the spacing of thermal contraction polygons is typically about five times the annual thermal skin depth (the depth to which the summer–winter thermal wave penetrates into the surface of a planetary body)⁴. Using reasonable values for the thermal diffusivity of N₂ ice⁵, we compute the annual thermal skin depth to be about 100 m, corresponding to thermal contraction polygons around 500 m across; this is nearly two orders of magnitude smaller than the observed polygons. Furthermore, contractional polygons require brittle failure of

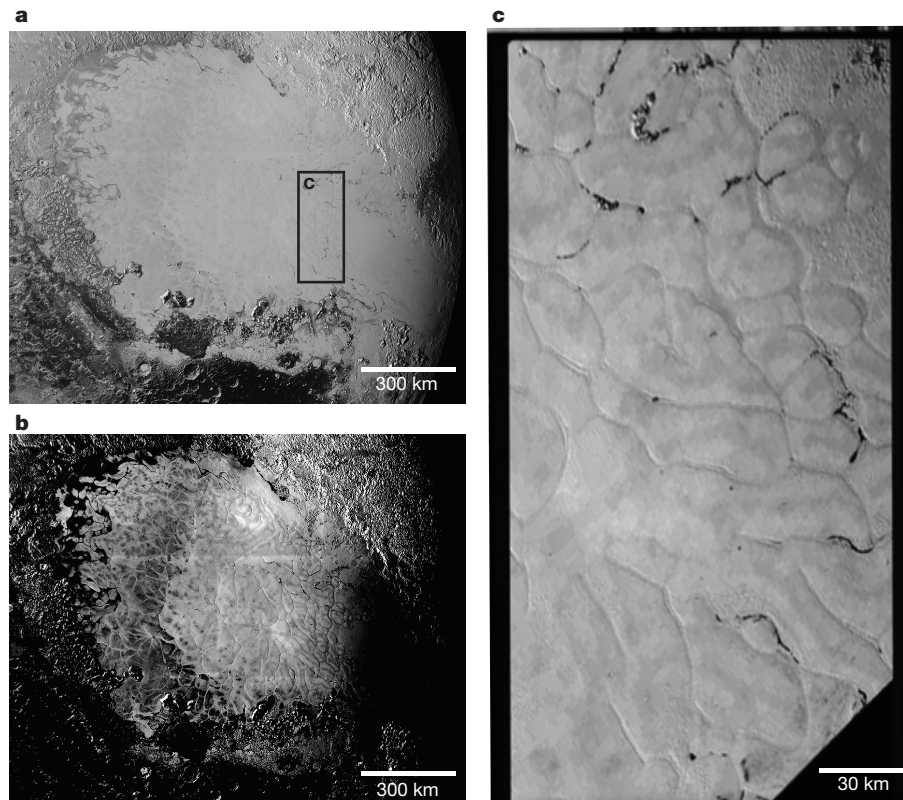


Figure 1 | New Horizon's image of Sputnik Planum on Pluto. A mosaic image of Sputnik Planum is shown in **a**. Within the centre of the ice field, where the ice is presumably thickest, the polygons are approximately 30 km across¹. Close to the edge, the average polygon diameter decreases to 20 km and then vanishes, leaving a smooth surface. A contrast-

enhanced version of **a** is given in **b** to better illuminate the polygons. The 'floating mountains' are observable within the edges of these polygons, and can be seen in **c**, the zoom of the rectangle in **a**. Image credit: NASA/John Hopkins University-Applied Physics Laboratory/Southwest Research Institute (2015).

¹Department of Earth, Atmospheric, and Planetary Sciences, Purdue University, West Lafayette, Indiana 47907, USA. ²Department of Physics and Astronomy, Purdue University, West Lafayette, Indiana 47907, USA.

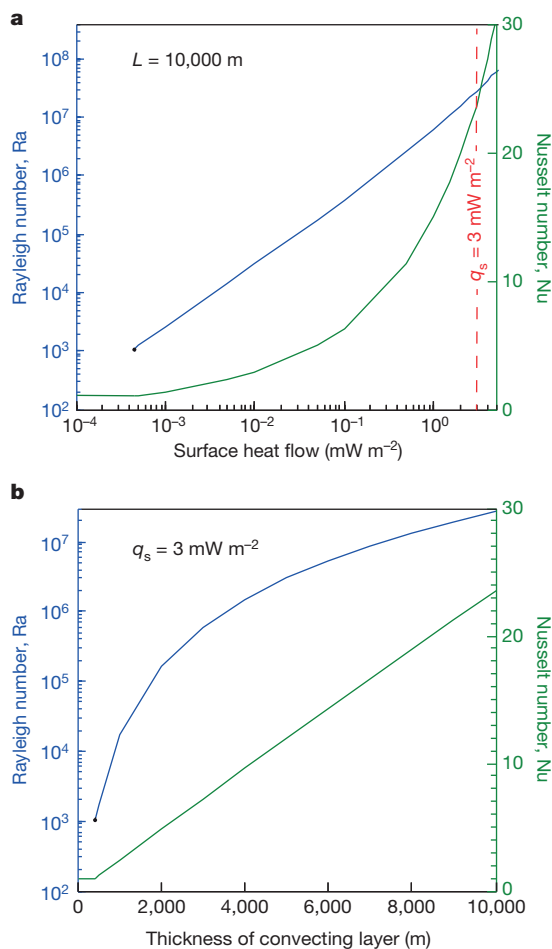


Figure 2 | Calculated convection for Sputnik Planum polygons. The calculated Rayleigh number and Nusselt number as a function of surface heat flow (q_s) and thickness of convecting layer (L) are shown in **a** and **b**, respectively. The blue line is the Rayleigh number; the green line is the Nusselt number (see Methods section). The black dot marks the point where the Rayleigh number reaches the critical value ($\sim 1,000$) at which convection just begins. At this point, the Nusselt number equals 1, and heat is transferred entirely by conduction. The calculation as a function of surface heat flow (**a**) shows that the Rayleigh number remains above $\sim 1,000$ for surface heat flows down to $4 \times 10^{-4} \text{ mW m}^{-2}$ for a 10-km-thick N_2 layer. The estimated surface heat flow for Pluto (3 mW m^{-2}) is marked by a vertical red line. For a constant heat flow of 3 mW m^{-2} , the Rayleigh number (**b**) decreases with the thickness of the convecting layer until convection stops at a thickness of 425 m.

the ice. However, viscoelastic deformation of N_2 ice over annual (248 Earth years) and diurnal (153 Earth hours) cycles⁶ on Pluto can easily relax differential stresses on this timescale, preventing brittle failure in response to such slowly building stress (the Maxwell time, over which stresses relax by $1/e$, of N_2 ice at 40 K is about 4 min at a stress of 0.1 MPa). Although there are other ways to generate polygons (such as compaction of sediments over heavily cratered terrain⁷, extensional tectonic processes⁸ or contraction of cooling cryovolcanic flows^{9,10}, these processes occur on timescales that are significantly longer than the Maxwell time, and are inconsistent with the lack of craters and observed flow features within this region.

The viability of Rayleigh–Bénard convection as an explanation for Pluto’s polygons depends critically on the thickness of the convecting layer¹¹. Because spectral data only probes micrometres into the surface, the N_2 ice could be only a thin surface veneer, making convection impossible. However, this possibility is unlikely given the high exchange rates of the N_2 atmosphere with a surface ice reservoir¹². We can estimate the ice thickness from the observed polygon size

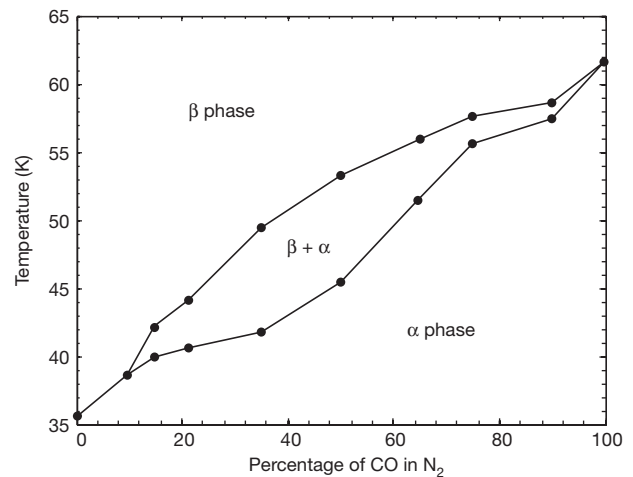


Figure 3 | The N_2 –CO phase diagram. We obtained this figure by collecting data points from experimental measurements¹⁷. Above CO concentrations of 10%, both phases of nitrogen are stable at Pluto’s surface temperatures.

and the depth to diameter ratio (that is, aspect ratio) for a Rayleigh–Bénard convection cell. Laboratory convection experiments and three-dimensional numerical modelling almost invariably predict aspect ratios near 3:1 (ref. 11), as assumed here. Some two-dimensional numerical simulations of convection in fluids with strongly temperature-dependent viscosity predict larger aspect ratios^{13,14}, implying a thinner ice layer. However, N_2 -ice viscosity depends only weakly on temperature¹⁵ and falls in the small-viscosity-contrast regime that precludes aspect ratios larger than 3:1 (refs 14, 16). Moreover, if Pluto’s ‘floating mountains’ (see Fig. 1c) are truly supported by their buoyancy, then their heights, widths and the small density contrast between N_2 ice and water ice requires an N_2 -ice thickness of at least 5 km, as shown in the Methods. We account for uncertainty in the layer thickness by widely varying the depth of the convection cell within our model (Fig. 2).

Using extrapolated N_2 -ice rheology measurements¹⁵ (see Methods and parameters used in equation (9)), a surface temperature of 33 K, and a surface heat flux q_s of $\sim 3 \text{ mW m}^{-2}$ (consistent with the radiogenic heat generated by a carbonaceous chondrite core about 900 km in radius, as suggested by Pluto’s mean density), we calculate a Rayleigh number $Ra > 10^6$ and an interior temperature of approximately 40 K for the N_2 -ice layer. This Rayleigh number is four orders of magnitude greater than the critical value that denotes the onset of convection ($Ra_{\text{crit}} \approx 1,000$), suggesting that Sputnik Planum’s N_2 ice is vigorously convecting.

Figure 2 shows the calculated Rayleigh and Nusselt numbers for a range of surface heat flows and N_2 thicknesses. In our model, the Rayleigh number remains above the critical value for surface heat flows as small as $4 \times 10^{-4} \text{ mW m}^{-2}$ (see Fig. 2a), suggesting that our results are robust against uncertainties in our estimated surface heat flux. Figure 2b illustrates that as the thickness of the convecting cell decreases, the Rayleigh number drops until convection ceases at a thickness of 425 m (at a nominal heat flow of 3 mW m^{-2}). Thus, the observed decrease in polygon size away from the centre of Sputnik Planum, and their absence at its edges (see Fig. 1), both suggest that the depth of Sputnik Planum’s N_2 ice is thickest at the centre, and thins to around 400 m near the edges, where the polygons are absent. This result is consistent with the hypothesis that N_2 ice in Sputnik Planum fills a topographic basin.

From the calculated average velocity of convection, $\sim 1.5 \text{ cm yr}^{-1}$ (the equation for velocity is given in the Methods), we compute the time needed for the ice surface to renew itself, and therefore the maximum age of the surface of Sputnik Planum, to be about one million years. This is consistent with the lack of significant cratering, and further constrains the existing age estimates of a few hundred

million years¹ by two orders of magnitude. The convection model also correctly predicts the topography of the polygons. The centres of convection cells are underlain by (relatively) warm rising currents and should therefore stand higher than the edges of the cells, where cooler ice descends. The elevation of the polygon centres above their edges is estimated from the convective buoyancy stress to be about 80 m (explained in more detail in the Methods), in good agreement with observations¹.

Our predicted central temperature of the convection cell of 40 K is close to the α -to- β phase transition temperature for N₂ ice ($T_{\alpha\beta}$ = 35.61 K for pure N₂ ice, but is higher if CO is dissolved in the N₂ ice¹⁷). Furthermore, high concentrations of CO have been reported on Sputnik Planum¹. CO and N₂ form a complete solid solution series (see Fig. 3). With a CO concentration >10%, both phases are stable at 40 K (ref. 17), which may allow a phase-change-induced mixed convection system to develop at these high Rayleigh numbers¹⁸. This may consist of warm β N₂ upwelling in the convection cells, while α N₂ concentrates in the troughs of the polygonal features, which might explain the albedo difference between the polygons and troughs. Owing to the difference in absorption spectra for each phase, New Horizon's Alice instrument can test this prediction of our convection model. Alternatively, a two-layer convection system may develop, in which the β N₂ forms the lower, deeper convection cell while α N₂ forms a layer of convection cells above. However, the positive Clapeyron slope of the α -to- β phase transition makes this alternative scenario unlikely¹⁹. Rather, the exothermic β -to- α phase change encourages single-cell overturn and produces a more stable convecting regime²⁰.

The N₂-ice mass in Sputnik Planum seems to be the largest concentration of N₂ on Pluto. If our estimated 10-km ice thickness were evenly spread across the planet it would form a layer about 350 m thick that would, if converted to vapour, produce an atmosphere with a surface pressure of about one bar instead of the currently observed pressure of approximately ten microbars. The present atmospheric pressure is presumably controlled by the vapour pressure of N₂ ice at the current surface temperature, so that N₂ can either evaporate or condense onto the ice reservoir in Sputnik Planum as temperature varies during Pluto's seasons and annual excursions from the sun. N₂ ice should thus be mobile across Pluto's surface. We do not at present understand why most of the N₂ on Pluto is concentrated in what appears to be the basin of a large ancient impact crater: that must be the subject of future climatological studies. However, it is an observational fact that most of the N₂ on Pluto is concentrated in a single large mass that lies in a basin nearly at the equator rather than at its poles, which is perhaps related to Pluto's large obliquity. The polygonal surface features of this mass indicate that it is vigorously convecting, and this convection is driven by the small amount of heat conducted through Pluto's lithosphere.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 27 November 2015; accepted 11 April 2016.

1. Stern, S. *et al.* The Pluto system: initial results from its exploration by New Horizons. *Science* **350**, <http://dx.doi.org/10.1126/science.aad1815> (2015).
2. Harry, D. & Gozdzik, J. Ice wedges: growth, thaw transformation, and palaeoenvironmental significance. *J. Quat. Sci.* **3**, 39–55 (1988).
3. Kindle, E. Some factors affecting the development of mud-cracks. *J. Geol.* **25**, 135–144 (1917).
4. Lachenbruch, A. Mechanics of thermal contraction cracks and ice-wedge polygons in permafrost. *Geol. Soc. Am. Spec. Pap.* **70**, 1–66 (1962).
5. Stachowiak, P., Sumarokov, V., Mucha, J. & Jeżowski, A. Thermal conductivity of solid nitrogen. *Phys. Rev. B* **50**, 543–546 (1994).
6. Hansen, C. & Paige, D. Seasonal nitrogen cycles on Pluto. *Icarus* **120**, 247–265 (1996).
7. McGill, G. & Hills, L. Origin of giant Martian polygons. *J. Geophys. Res.* **97**, 2633–2647 (1992).
8. Pechmann, J. The origin of polygonal troughs on the Northern Plains of Mars. *Icarus* **42**, 185–210 (1980).
9. Freed, A. *et al.* On the origin of graben and ridges within and near volcanically buried craters and basins in Mercury's northern plains. *J. Geophys. Res.* **117**, E00L06 (2012).
10. Blair, D. *et al.* The origin of graben and ridges in Rachmaninoff, Raditladi, and Mozart basins, Mercury. *J. Geophys. Res. Planets* **118**, 47–58 (2013).
11. Schubert, G., Turcotte, D. & Olson, P. *Mantle Convection in the Earth and Planets* (Cambridge Univ. Press, 2001).
12. Stern, S., Porter, S. & Zangari, A. On the roles of escape erosion and the viscous relaxation of craters on Pluto. *Icarus* **250**, 287–293 (2015).
13. Barr, A. & Hammond, N. A common origin for ridge-and-trough terrain on icy satellites by sluggish lid convection. *Phys. Earth Planet. Inter.* **249**, 18–27 (2015).
14. Kameyama, M. & Ogawa, M. Transitions in thermal convection with strongly temperature-dependent viscosity in a wide box. *Earth Planet. Sci. Lett.* **180**, 355–367 (2000).
15. Yamashita, Y., Kato, M. & Arakawa, M. Experimental study on the rheological properties of polycrystalline solid nitrogen and methane: implications for tectonic processes on Triton. *Icarus* **207**, 972–977 (2010).
16. Moresi, L. & Solomatov, V. Numerical investigation of 2D convection with extremely large viscosity variations. *Phys. Fluids* **7**, 2154–2162 (1995).
17. Angwin, M. Nitrogen-carbon monoxide phase diagram. *J. Chem. Phys.* **44**, 417–418 (1966).
18. Zhao, W., Yuen, D. & Honda, S. Multiple phase transitions and the style of mantle convection. *Phys. Earth Planet. Inter.* **72**, 185–210 (1992).
19. Christensen, U. & Yuen, D. The interaction of a subducting lithospheric slab with a chemical or phase boundary. *J. Geophys. Res.* **89**, 4389–4402 (1984).
20. Schubert, G., Yuen, D. & Turcotte, D. Role of phase transitions in a dynamic mantle. *Geophys. J. Int.* **42**, 705–735 (1975).

Acknowledgements We thank all of the New Horizons team members, without whom none of this work would have been possible. We also thank T. Bowling, D. Minton, B. Hogan, J. Kendall, B. Link and C. Milbury for discussions. A.J.T. thanks the Fredrick N. Andrews Fellowship for funding.

Author Contributions A.J.T. and H.J.M. conceived this work, developed the parameterized convection model, and conducted Rayleigh number calculations for this paper. J.K.S. developed Maxwell time arguments for ruling out thermal contraction, computed the surface and subsurface temperatures of Pluto, and calculated atmospheric pressures. A.M.F. advised A.J.T., and helped to edit and revise the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to A.J.T. (atrowbr@purdue.edu).

Reviewer Information Nature thanks G. Schubert and the other anonymous reviewer(s) for their contribution to the peer review of this work.

METHODS

Maxwell time calculation. The Maxwell time (τ_m in units of seconds) is determined from the ratio of the effective viscosity (η_{eff}) to the shear modulus (μ in units of pascals):

$$\tau_m = \frac{\eta_{\text{eff}}}{\mu} \quad (1)$$

The shear modulus was determined from experimentally measured shear wave velocities (v_s in units of metres per second)²¹, while the viscosity parameters are derived from the measurements of N_2 ice¹⁵ and are listed below. The shear modulus is related to the shear wave velocity and material density (ρ in units of kilograms per cubic metre) by the expression:

$$\mu = v_s^2 \rho \quad (2)$$

At a differential stress of 0.1 MPa and a temperature of 30 K, the Maxwell time is about 4 min.

Temperature-dependent parameters for Rayleigh number. The temperature dependent values for the parameters in equation (7) (see below) were determined from best-fitting experimental data²¹. The density, volume coefficient of expansion (α in units of per kelvin), and thermal conductivity (k in units of watts per millikelvin) of N_2 are expressed by the following equations:

$$\rho = 0.0134T^2 - 0.6981T + 1038.1 \quad (3)$$

$$\alpha = (2 \times 10^{-6})T^2 - 0.0002T + 0.006 \quad (4)$$

$$k = 0.1802T^{0.1041} \quad (5)$$

The thermal diffusivity (κ in units of metres squared per second) was determined from the relationship $\kappa = k/(\rho c_p)$, where c_p (in units of joules per kilogram per kelvin) is the specific heat at constant pressure. Specific heat is also temperature dependent, which is expressed as:

$$c_p = 926.91e^{0.0093T} \quad (6)$$

Parameterized convection model. The Rayleigh number assesses the vigour of thermal convection in a fluid layer under the influence of gravity and a downward-increasing thermal gradient. The Rayleigh number is essentially the ratio between the timescales for conductive cooling and buoyancy-driven overturn of the viscous fluid. The conventional expression for the Rayleigh number¹¹ requires knowledge of the temperature difference between the top and bottom of the convecting layer. Because the thermal gradient in Pluto is a priori unknown, but the surface heat flow can be estimated at about 3 mW m^{-2} from internal heat production, we employ a version of the Rayleigh number based on surface heat flow¹¹:

$$Ra_q = \frac{\alpha \rho g L^4}{\kappa \eta} q_s \quad (7)$$

where α is the volume coefficient of thermal expansion, ρ is the fluid density, g is the acceleration of gravity (0.62 m s^{-2} on Pluto), L is the depth of the convecting layer, κ is thermal diffusivity, k the thermal conductivity, η is the viscosity (which depends strongly on temperature and stress in N_2 ice), and q_s is the surface heat flow. For Sputnik Planum, we used (see below) measured, temperature-dependent values for N_2 in equation (1). This version of the Rayleigh number is related to the more standard version Ra through the Nusselt number Nu as follows: $Ra_q = Nu \times Ra$. The Nusselt number is the ratio between the total heat transported by both convection and conduction to conductive heat transport only and is given by:

$$Nu = \left(\frac{Ra}{Ra_{\text{crit}}} \right)^\beta \text{ for } Ra > Ra_{\text{crit}} \quad (8)$$

The critical Rayleigh number Ra_{crit} is of order 10^3 , depending on detailed boundary conditions, and β has been measured to be 0.31 over a wide range of Ra values²².

Equations (7) and (8) together define the “parameterized convection model”, which has been widely used to model heat transport in planetary mantles^{11,23}. The most important variable in these models is the viscosity of the convecting fluid. Unlike ideal liquids, the viscosity of a hot, creeping solid is a sensitive function of both deviatoric stress σ and temperature T and is often parameterized by the form²⁴:

$$\eta_{\text{eff}} = \frac{\sigma^{1-n}}{3A} e^{Q/RT} \quad (9)$$

where Q is the activation enthalpy for creep and R is the gas constant. A is a constant that, along with Q , and n , must be determined experimentally. The viscosity parameters we determined for N_2 ice used in equation (9) are activation energy $Q = 3.5 \text{ kJ mol}^{-1}$, $n = 2.2$ and $A = 3.5 \times 10^{-12} \text{ Pa}^{-n} \text{ s}^{-1}$.

Because the effective viscosity depends strongly on both the temperature and stress in the convecting system, which vary widely from place to place, it is important to understand how to define them in a meaningful way. Previous work¹¹ has shown that the best choice is to use the mean temperature and buoyancy stress for accurate estimates of convective vigour, a choice that we follow here.

As they stand, equations (7) to (9) do not define a closed system and more information is required to compute Ra , even given the heat flow and material properties of the convecting fluid. The system can, however, be closed by recognizing that the mean temperature in a convecting fluid is determined by the surface temperature, T_s , heat flow q_s and the temperature drop across the cold (surface) boundary layer, whose thickness is itself determined from the Nusselt number. We thus set:

$$T = T_s + \frac{q_s L}{k(Nu + 1)} \quad (10)$$

where we have ignored the adiabatic increase of temperature in the convecting layer, a valid approximation for a thin layer, such as the N_2 -ice deposit in Sputnik Planum. Further adding an equation for the average deviatoric stress in convecting plumes:

$$\sigma = \left(\frac{Ra_q}{Nu^2} \right) \frac{\eta_{\text{eff}} \kappa}{L^2} \quad (11)$$

Equations (7) to (11) now define a closed, if highly nonlinear, system that can readily be solved by numerical methods to define most of the properties of the convecting layer from the properties of the fluid, the surface heat flow and surface temperature.

Viscosity of N_2 ice. Equations (3) to (6) describe all the material parameters in equation (7) except for viscosity. The stress-dependent parameters, A and n , within the viscosity equation are directly quoted from previous works¹⁵. The temperature-dependent parameter was determined from existing stress and strain rate (that is, viscosity) measurements for solid N_2 at 45 K and 56 K (ref. 15). By matching data points for the viscosity measured at two temperatures under the same applied strain rate, we can solve for the temperature-dependent parameter, Q , in equation (9).

Velocity of the convecting fluid. The mean velocity of a convecting layer is computed by comparing the surface heat flow to the rate at which warm fluid moves towards the surface and deposits its thermal energy. This equality can be written in terms of the Nusselt number Nu as:

$$\bar{v}_{\text{conv}} = \frac{\kappa}{L} (Nu - 1) \quad (12)$$

where κ is the thermal diffusivity and L is the depth of the convective cell.

Topographic relief of convecting terrain. We estimate the difference in elevation h between the upwelling centres of the polygons and their sinking margins by equating the buoyancy stress in the convecting fluid to the stress generated by topography, ρgh , where ρ is the density and g is Pluto's surface acceleration of gravity (0.62 m s^{-2}). The buoyancy stress is equal to the density deficit of the warm, rising fluid, $\rho \alpha \Delta T$, where α is the volume coefficient of expansion and ΔT is the temperature difference between the hot and cold boundaries of the convecting system. The density deficit is multiplied by the height of the convection cell, L , times g to define the convective stress, from which we deduce:

$$h = \alpha \Delta T L \quad (13)$$

However, ΔT is not known a priori. We can define it more precisely in terms of quantities better defined in convecting systems by exploiting the conventional definition of the Rayleigh number Ra to solve for ΔT and write:

$$h = \frac{Ra}{\rho g} \frac{\kappa \eta_{\text{eff}}}{L^2} \quad (14)$$

Inserting this expression into our system of parameterized convection equations for the nominal case of a heat flow of 3 mW m^{-2} and $L = 10 \text{ km}$ yields an estimate of about 80 m for the difference in elevation between the centre and edges of the Rayleigh–Bénard cells, as we report in the text.

The ‘floating mountains’ of Sputnik Planum. Dark material congregates along the edges of the Sputnik Planum polygons (see Fig. 1). These hills (currently called ‘floating mountains’ in NASA press releases) rise hundreds of metres¹ above the surrounding terrain within Sputnik Planum. Owing to the albedo contrast with the surrounding N_2 ice, the material seems likely to be composed of water ice¹. Nearly

all of the ice chunks are located at the edges of the cells rather than the centres, and are arranged in lines and arcs that do not resemble the rims of submerged impact craters. Because such a non-random distribution of mountains is unlikely, we conclude that the mountains are afloat and are moved by N_2 convection to the edges of the polygons. Although it is possible that the downwelling limbs of the convection cells have aligned with grounded water-ice mountains in thinner N_2 ice, the polygonal arrangement of these mountains at the same distance scale as the mountain-free polygons strongly suggest that their arrangement was determined by the dynamics of the convection cells rather than vice versa.

If these ‘floating mountains’ are icebergs, then we can calculate the minimum depth of N_2 ice beneath each one that is needed to generate strong enough buoyancy forces to keep it afloat. According to Archimedes’ principle, the depth of the bottom of an iceberg of height h above the surface is:

$$d = \frac{h}{\left(\frac{\rho_n}{\rho_w} - 1\right)} \quad (15)$$

where ρ_n is the density of N_2 and ρ_w is the density of H_2O . At Pluto temperatures of ~ 37 K (ref. 1), the density of water ice is $\sim 930 \text{ kg m}^{-3}$ (refs 25, 26), and the density of N_2 is $\sim 1,030 \text{ kg m}^{-3}$ (from equation (3)). Using these densities and a height of 500 m for the iceberg topography, we calculate a minimum depth of 5 km.

The horizontal extent of these mountains also gives clues to their depths because tall, narrow cylindrical masses of ice are not stable: they would tilt to achieve a minimum gravitational energy configuration. The largest observed masses are about 5 km across (see Fig. 1), suggesting a minimum N_2 ice depth comparable to, or greater, than this distance.

Effect of r_η on the aspect ratio for Rayleigh–Bénard convection. The temperature-dependence parameter r_η is a non-dimensional ratio between the viscosities at the top and bottom of the convection cell that determines the regime of convection (transitional mode, stagnant-lid mode or small-viscosity-contrast mode) as well as

the aspect ratio¹⁴. We show below that the N_2 -ice layer in Sputnik Planum is well within the small-viscosity-contrast regime, so special considerations for convection in strongly temperature-dependent fluids do not apply.

The ratio r_η is given by the following formula¹⁴:

$$r_\eta = e^{E(T_b - T_s)} \quad (16)$$

where T_b is the temperature at the bottom of the convection and T_s is the surface temperature. Within equation (16), the E term is a constant determined by fitting the temperature-dependent viscosity equation to rheologic measurements for a material¹⁵ and is defined as:

$$E = \frac{Q}{RT^2} \quad (17)$$

where T is the mean temperature, R is the gas constant, and Q is the activation energy (see equation (9)). Using a temperature of 45 K and the activation energy for N_2 (see parameters used in equation (9)), we calculated an E constant of ~ 0.2 , corresponding to an r_η value of ~ 15 for N_2 . This value for r_η places the N_2 in Sputnik Planum within the small-viscosity-contrast convection regime, where an aspect ratio greater than 3:1 is not possible for $Ra > 10^6$ (refs 14, 16).

21. Scott, T. Solid and liquid nitrogen. *Phys. Rep.* **27**, 89–157 (1976).
22. Niemela, J., Skrbek, L., Sreenivasan, K. & Donnelly, R. Turbulent convection at very high Rayleigh numbers. *Nature* **404**, 837–840 (2000).
23. Schubert, G. Numerical models of mantle convection. *Annu. Rev. Fluid Mech.* **24**, 359–394 (1992).
24. Karato, S. *Deformation of Earth Materials: an Introduction to the Rheology of Solid Earth* 338–362 (Cambridge Univ. Press, 2012).
25. Eisenberg, D. S. & Kauzmann, W. *The Structure and Properties of Water* 296 (Clarendon Press, 1969).
26. Hobbs, P. *Ice Physics* 346 (Oxford Univ. Press, 2010).

Convection in a volatile nitrogen–ice–rich layer drives Pluto’s geological vigour

William B. McKinnon¹, Francis Nimmo², Teresa Wong¹, Paul M. Schenk³, Oliver L. White⁴, J. H. Roberts⁵, J. M. Moore⁴, J. R. Spencer⁶, A. D. Howard⁷, O. M. Umurhan⁴, S. A. Stern⁶, H. A. Weaver⁵, C. B. Olkin⁶, L. A. Young⁶, K. E. Smith⁴ & the New Horizons Geology, Geophysics and Imaging Theme Team*

The vast, deep, volatile-ice-filled basin informally named Sputnik Planum is central to Pluto’s vigorous geological activity^{1,2}. Composed of molecular nitrogen, methane, and carbon monoxide ices³, but dominated by nitrogen ice, this layer is organized into cells or polygons, typically about 10 to 40 kilometres across, that resemble the surface manifestation of solid-state convection^{1,2}. Here we report, on the basis of available rheological measurements⁴, that solid layers of nitrogen ice with a thickness in excess of about one kilometre should undergo convection for estimated present-day heat-flow conditions on Pluto. More importantly, we show numerically that convective overturn in a several-kilometre-thick layer of solid nitrogen can explain the great lateral width of the cells. The temperature dependence of nitrogen-ice viscosity implies that the ice layer convects in the so-called sluggish lid regime⁵, a unique convective mode not previously definitively observed in the Solar System. Average surface horizontal velocities of a few centimetres a year imply surface transport or renewal times of about 500,000 years, well under the ten-million-year upper-limit crater retention age for Sputnik Planum². Similar convective surface renewal may also occur on other dwarf planets in the Kuiper belt, which may help to explain the high albedos shown by some of these bodies.

Sputnik Planum (SP) is the most prominent geological feature on Pluto revealed by NASA’s New Horizons mission. It is a ~900,000 km² oval-shaped unit of high-albedo plains (Fig. 1a) set within a topographic basin at least 2–3 km deep (Fig. 1b). The basin’s scale, depth and ellipticity (~1,300 × 1,000 km), and rugged surrounding mountains, suggest an origin as a huge impact—one of similar scale to its parent body as Hellas on Mars or South Pole–Aitken on the Moon⁶. The central and northern regions of SP display a distinct cellular/polygonal pattern (Fig. 1c). In the bright central portion, the cells are bounded by shallow troughs locally up to 100 m deep (Fig. 1d), and the centres of at least some cells are elevated by ~50 m relative to their edges². The southern region and eastern margin of SP do not display cellular morphology, but instead show featureless plains and dense concentrations of kilometre-scale pits².

Impact craters have not been confirmed on SP either in New Horizons mapping at a scale of 350 m per pixel, or in high-resolution strips (resolutions as fine as 80 m per pixel). The crater retention age of SP is very young, no more than ~10 Myr based on models of the impact flux of small Kuiper belt objects onto Pluto⁷. This indicates renewal, burial or erosion of the surface on this timescale or shorter. Evidence for all three processes is seen in the form of possible convective overturn, glacial inflow of volatile ice from higher standing terrains at the eastern margin, and likely sublimation landforms such as the pits². In addition, the apparent flow lines around obstacles in northern SP and the pronounced distortion of some fields

of pits in southern SP are evidence for the lateral, advective flow of SP ices^{1,2}.

From New Horizons spectroscopic mapping, N₂, CH₄ and CO ice all concentrate within Sputnik Planum³. All three ices are mechanically weak, van der Waals bonded molecular solids and are not expected to be able to support appreciable surface topography over any great length of geological time^{4,8–10}, even at the present surface ice temperature of Pluto (37 K)¹. This is consistent with the overall smoothness of SP over hundreds of kilometres (Fig. 1b). Convective overturn that reaches the surface would also eliminate impact and other features, and below we estimate numerically the timescale for SP’s surface renewal.

Quantitative radiative transfer modelling of the relative surface abundances of N₂, CH₄ and CO ices within SP¹¹ shows that N₂ ice dominates CH₄ ice, especially in the central portion of the planum (the bright cellular plains) where the cellular structure is best defined topographically (Fig. 1d). Ices at depth need not match the surface composition, but continuous exposure (such as by convection) makes this more likely. N₂ and CO ice have nearly the same density (close to 1.0 g cm^{−3}), whereas CH₄ ice is half as dense as this². Hence water-ice blocks can float in solid N₂ or CO, but not in solid CH₄. Water ice has been identified in the rugged mountains that surround SP³, and blocks and other debris shed from the mountains at SP’s periphery appear to be floating²; moreover, glacial inflow appears to carry along water-ice blocks, and these blocks almost exclusively congregate at the margins of the cells/polygons, consistent with being dragged to the downwelling limbs of convective cells (Fig. 2a). This indicates that while CH₄ ice is present within SP, it is not likely to be volumetrically dominant. In terms of convection, we concentrate on the rheology of N₂ ice.

Deformation experiments for N₂ ice show mild power-law creep behaviour (strain rate proportional to stress to the $n = 2.2 \pm 0.2$ power) and a modest temperature dependence of its viscosity⁴. N₂ diffusion creep ($n = 1$) has also been predicted^{10,12}, but not yet observed experimentally. Convection in a layer occurs if the critical Rayleigh number (Ra_{cr}) is exceeded. The Rayleigh number, the dimensionless measure of the vigour of convection, for a power-law fluid heated from below is given by¹³

$$Ra = \frac{\rho g \alpha A^{1/n} \Delta T D^{(2+n)/n}}{\kappa^{1/n} \exp(E^*/nRT)} \quad (1)$$

where D is the thickness of the convecting layer, κ is the thermal diffusivity, g is the acceleration due to gravity, ρ the ice layer density, α the volume thermal expansivity, ΔT the superadiabatic temperature drop across the layer, and A is the pre-exponential constant in the relationship between stress and strain-rate, E^* is the activation energy of the dominant creep mechanism, and R is the gas constant.

¹Department of Earth and Planetary Sciences and McDonnell Center for the Space Sciences, Washington University in St Louis, Saint Louis, Missouri 63130, USA. ²Department of Earth and Planetary Sciences, University of California Santa Cruz, Santa Cruz, California 95064, USA. ³Lunar and Planetary Institute, Houston, Texas 77058, USA. ⁴National Aeronautics and Space Administration (NASA) Ames Research Center, Moffett Field, California 94035, USA. ⁵Johns Hopkins University Applied Physics Laboratory, Laurel, Maryland 20723, USA. ⁶Southwest Research Institute, Boulder, Colorado 80302, USA. ⁷Department of Environmental Sciences, University of Virginia, Charlottesville, Virginia 22904, USA.

*A list of authors and affiliations appears at the end of the paper.

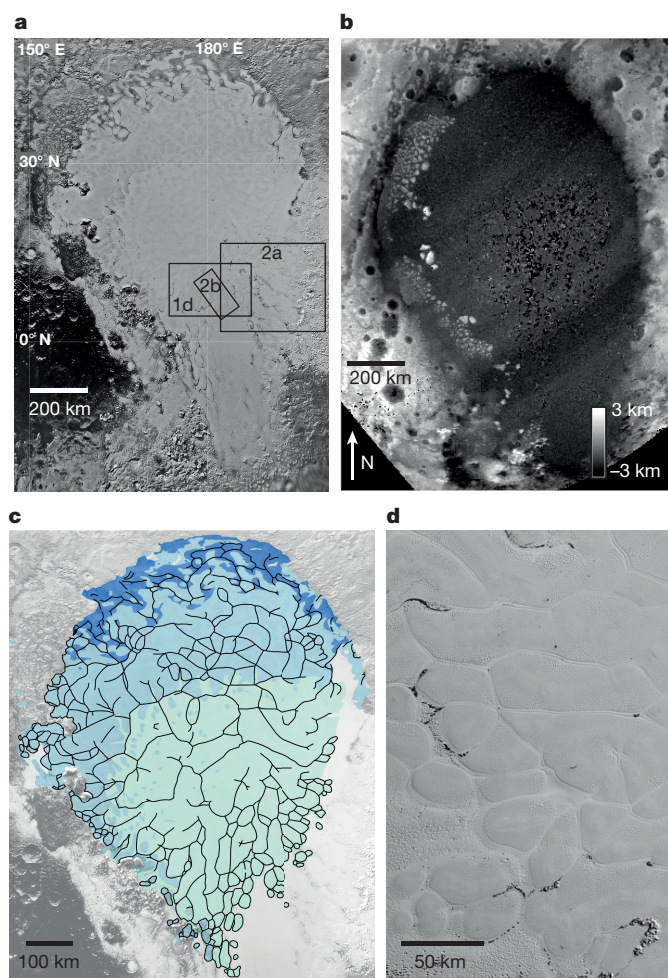


Figure 1 | Image, topographic and map views of Sputnik Planum, Pluto. **a**, Base map showing locations of some figure panels. **b**, Stereo-derived topography, showing that Sputnik Planum (SP) lies within a kilometres-deep basin (depth coded on greyscale, see key at bottom right). Southwest-northeast banding and central basin 'speckle' are artefacts or noise (Methods); elevations are relative. **c**, Map of troughs (black lines), which define cell boundaries (note enlarged scale compared with **a** and **b**). Cell size increases and/or becomes less well connected towards SP centre, consistent with a thickened N_2 ice layer there. Aquamarine shading indicates 'bright cellular plains', within which troughs are topographically defined. **d**, 350 m per pixel MVIC image (position shown in **a**) that shows cellular/polygonal detail (north is to right).

The critical Rayleigh number depends on the temperature drop and the associated change in viscosity¹³, as deformation mechanisms are thermally activated processes. For a given ΔT , the Ra_{cr} implies a critical or minimum layer thickness, D_{cr} , below which convection cannot occur. This is illustrated in Fig. 3 for N_2 ice. We assume an average ice surface temperature of 36 K set by vapour-pressure equilibrium over an orbital cycle¹⁴, and an upper limit on the basal temperature set by the N_2 ice melting temperature of 63 K (ref. 15). From Fig. 3 we conclude that convection in solid nitrogen on Pluto is a facile process: critical thicknesses are generally low, less than 1 km, as long as the necessary temperatures at depth are achieved.

The temperature profile in the absence of convection is determined by conduction. N_2 ice has a low thermal conductivity¹⁵, which together with a present-day radiogenic heat flux for Pluto of roughly 3 mW m^{-2} implies a conductive temperature gradient of $\sim 15 \text{ K km}^{-1}$. Over Pluto's history, radiogenic heat has dominated Pluto's internal energy budget^{16,17}; we argue that relatively unfractionated, solar composition carbonaceous chondrite is the best model for the rock component of worlds accreted in the cold, distant regions of the Solar

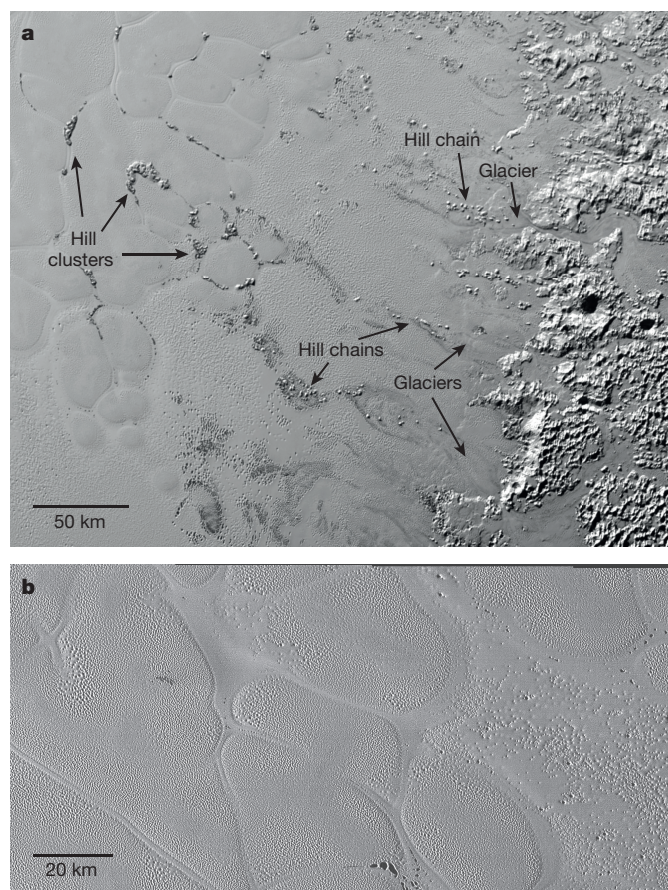


Figure 2 | High-resolution images of cellular terrain within SP.

a, Kilometre-scale hills appear to emanate from uplands to the east (at right), and are probably darker water-ice blocks and methane-rich debris (arrows) that have broken away and are being carried by denser, N_2 -ice-dominated glaciers into SP, where they become subject to the convective motions of SP ice, and are pushed to the downwelling edges of the cells at left. **b**, Part of the highest-resolution image sequence taken by New Horizons (80 m per pixel); surface texture (for example, pitting) concentrates towards cell boundaries and in regions apparently unaffected by convection (such as at right, see text).

System¹⁶. The abundances of U, Th, and ^{40}K are consistent across the most primitive individual examples of this meteorite group (the CI chondrites), to within 15% (ref. 18), and Pluto's density implies that about 2/3 of its mass could be composed of solar composition rock (the rest being ices and carbonaceous material)¹⁹. Nevertheless, regional and temporal variations in heat flow are possible, so Fig. 3 illustrates the temperatures reached as a function of depth, with the conclusion being that even under broad variations in heat flow, temperatures sufficient to drive convection in SP are plausible for N_2 -ice layers thicker than $\sim 500 \text{ m}$.

Clearly, the horizontal scale of the cells in SP (Figs 1d, 2a, b) should reflect the vertical scale (depth) of the SP basin ice fill, but this presents a problem. For isoviscous Rayleigh–Bénard convection, the aspect ratio (width/depth) of well-developed convection cells is near unity. Numerical calculations by us for Newtonian and non-Newtonian convection in very wide 2D domains, but without temperature-dependent viscosity, give aspect ratios near 1 (Methods). If the cells/polygons on Sputnik Planum are the surface expression of convective cells, then cell diameters (wavelengths λ) of 20–40 km imply depths to the base of the N_2 ice layer in SP of about 10–20 km. This is very deep, and much deeper than any likely impact basin, especially as the surface of SP is already at least 2–3 km below the surrounding terrain (Fig. 1b). The deepest impact basins of comparable scale known on any major icy world are on Iapetus, a body of much lower density (hence lower

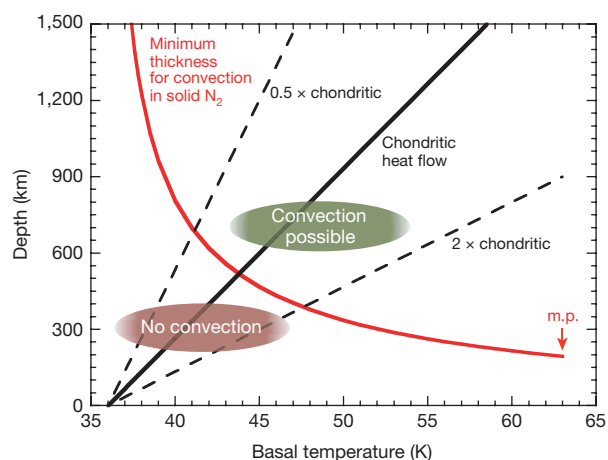


Figure 3 | Minimum thickness for convection in a layer of solid N_2 ice on Pluto, as a function of basal temperature. Convection can occur above the solid red curve provided a sufficient perturbation exists (area labelled 'convection possible'). Limit is based on numerical and laboratory experiments and theory and creep measurements for nitrogen ice (Methods). Basal temperatures due to conductive heat flow (3 mW m^{-2}) from Pluto are shown for comparison (solid black line), along with variations of a factor of 2 in heat flow (dashed black lines). For approximately present-day chondritic heat flows, basal temperatures exceed the convective threshold for layer thicknesses in excess of about 500 m. In contrast, the minimum thickness for convection by volume diffusion creep would plot off the graph to the upper right.

rock abundance and heat flow) and surface gravity than Pluto. Gravity scaling the depth from basins on Iapetus²⁰, we estimate the SP basin was initially no deeper than ~ 10 km total (that is, before filling by volatile ices or any isostatic adjustment).

The solution to this apparent problem (the SP ice thickness overestimate) is probably the temperature dependence of the N_2 ice viscosity. Given that the maximum ΔT across the SP N_2 layer is 27 K, the maximum corresponding Arrhenius viscosity ratio ($\Delta\eta$) for the experimentally constrained activation energy is ~ 150 (Methods); if we adopt the (larger) activation energy for volume diffusion²¹, this ratio potentially increases to $\sim 2 \times 10^5$. This potential range in $\Delta\eta$ strongly suggests that SP convects in the sluggish lid regime^{5,13,22}. In sluggish lid convection the surface is in motion and transports heat, but moves at a much slower pace than the deeper, warmer subsurface. A defining characteristic of this regime — depending on Ra_b (the Rayleigh number defined with the basal viscosity) and $\Delta\eta$ — is convection cells with large aspect ratios. This differs from isoviscous convection in which the aspect ratios are closer to one, or at the other end of the viscosity contrast spectrum, stagnant lid convection, in which aspect ratios are again closer to one but confined ('hidden') beneath an immobile, high-viscosity surface layer.

We illustrate such temperature-dependent viscosity convection numerically, using the finite element code CitCom²² (a typical example is shown in Fig. 4). Given that N_2 -ice rheology is imprecisely known (unlike well-studied geological materials such as olivine or water ice), we survey different combinations of Ra_b and $\Delta\eta$ in a Newtonian framework (similar to previous work^{5,22}), but with a rigid (no-slip) lower boundary condition appropriate to the SP ice layer (Methods). We find that aspect ratios easily reach values of 2 or 3 (or λ/D of 4 or 6), regardless of initial perturbation wavelength. In such instances cell dimensions between 20 km and 40 km across could imply a layer thickness as small as ~ 3 –6 km. We note that while these depths are not excessive, they are deep enough to carry buoyant, kilometre-scale water ice blocks. In addition, simulations with a free-slip lower boundary, which would apply to SP ice that is at or near melting at its base, yield aspect ratios as great as ~ 6 ($\lambda/D \approx 12$).

Numerical simulations can be tested against SP observations by assuming reasonable heat flows (say, chondritic within $\pm 50\%$) and

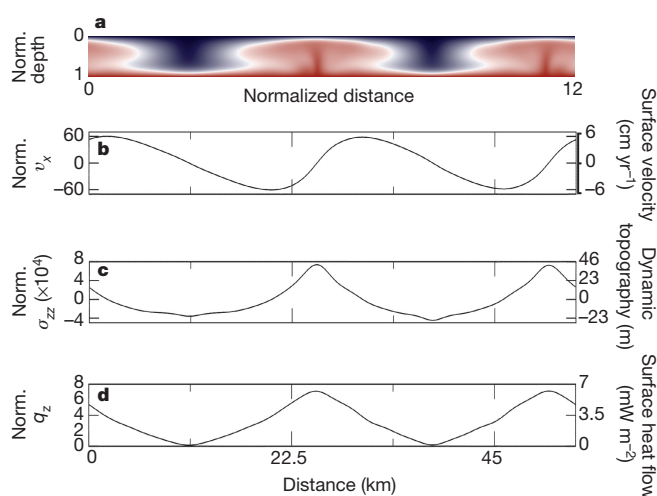


Figure 4 | Example numerical model of N_2 ice convection in SP.

a, Temperature field showing large-aspect-ratio plumes and downwellings. Basal Rayleigh number $Ra_b = 3 \times 10^5$, viscosity ratio $\Delta\eta = e^6 \approx 400$, Nusselt number (dimensionless heat flow) ≈ 3.2 . White contour denotes the median temperature. **b–d**, Corresponding horizontal surface velocities (**b**), surface normal stress and dynamic topography (**c**), and surface heat flows (**d**). Non-dimensional values are shown on the left, and dimensional values on the right assuming $D = 4.5$ km and $\Delta T = 20$ K. The calculated topography matches the scale seen within the bright cellular plains, and the average heat flow is consistent with radiogenic heat production in Pluto's rock fraction. Norm., normalized.

comparing the resulting dynamic topography with that observed. Non-dimensional surface horizontal velocity v_x , normal stress σ_{zz} , and heat flow q_z for the example calculation are shown in Fig. 4b, c. To dimensionalize we choose $D = 4.5$ km and $\Delta T = 20$ K to match the typical horizontal scale of the cells (for example, nearly 30 km, with a convective aspect ratio of 3) and give a chondritic heat flow (see Methods). The dynamic topography due to the thermal buoyancy of the flow is given by $\sigma_{zz}/\rho g$, and its scale is given at the right hand side of Fig. 4c. This dynamic topography is consistent with available measurements^{1,2}. Average surface velocities (Fig. 4b) in this example are a few centimetres per year, which for the horizontal scale of cells on SP translates into a timescale to transport surface ice from the centre of a given upwelling to the downwelling perimeter of $\sim 500,000$ years. This is well within the upper limit for the crater retention age for the planum, ~ 10 Myr (ref. 2). The surface heat flow variation is also notable, nearly double the mean over upwellings and close to zero over downwellings. This means that fine scale topography such as pitting or suncups driven by N_2 sublimation² will be much more stable towards cell/polygonal edges, as the N_2 ice there will be as cold and viscous as the surface to considerable depth, which is consistent with the observations of surface texture² (for example, Fig. 2b). We also find slight topographic dimples over downwellings in some of our calculations, which may be related to trough formation at cell edges (Fig. 2b). The troughs themselves, however, are likely to be finite amplitude topographic instabilities of the sort seen on icy satellites elsewhere²³, and are not captured by these convection calculations given that velocities normal to domain boundaries are set to zero.

Convection in a kilometres-thick N_2 layer within Pluto's SP basin thus emerges as a compelling explanation for the remarkable appearance of the planum surface (Fig. 1). Sputnik Planum covers 5% of Pluto's surface, so having an N_2 ice layer several kilometres deep is equivalent to a global layer ~ 200 –300 m thick. This is consistent with Pluto's possible total cosmochemical nitrogen inventory²⁴, especially as Pluto's atmospheric nitrogen escape rate is much lower than previously estimated²⁵. For Pluto, SP acts an enormous glacial catchment or drainage basin, the major topographic trap for Pluto's surficial, flowing N_2 ice. SP is essentially a vast, frozen sea, one in which convective turnover

(now, and even more vigorously in the past) continually refreshes the surface volatile ice inventory. A sealing, superficial lag of less volatile ices or darker tholins cannot develop²⁴, and the atmospheric cycle of volatile transport is maintained. Moreover, larger Kuiper belt objects are known to be systematically brighter (more reflective) than their smaller cousins in the Kuiper belt²⁶. Convective renewal of volatile ice surfaces, as in a basin or basins similar to SP, may be one way in which the dwarf planets of the Kuiper belt maintain their youthful appearance.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 21 December 2015; accepted 19 April 2016.

- Stern, S. A. *et al.* The Pluto system: initial results from its exploration by New Horizons. *Science* **350**, aad1815 (2015).
- Moore, J. M. *et al.* The geology of Pluto and Charon through the eyes of New Horizons. *Science* **351**, 1284–1293 (2016).
- Grundy, W. *et al.* Surface compositions across Pluto and Charon. *Science* **351**, aad9189 (2016).
- Yamashita, Y., Kato, M. & Arakawa, M. Experimental study on the rheological properties of polycrystalline solid nitrogen and methane: implications for tectonic processes on Triton. *Icarus* **207**, 972–977 (2010).
- Hammond, N. P. & Barr, A. C. Formation of Ganymede's grooved terrain by convection-driven resurfacing. *Icarus* **227**, 206–209 (2014).
- Schenk, P. M. *et al.* A large impact origin for Sputnik Planum and surrounding terrains, Pluto? *AAS/Div. Planet. Sci. Meeting* 47, abstr. 200.06 (2015).
- Greenstreet, S., Gladman, B. & McKinnon, W. B. Impact and cratering rates onto Pluto. *Icarus* **258**, 267–288 (2015).
- Moore, J. M. *et al.* Geology before Pluto: pre-encounter considerations. *Icarus* **246**, 65–81 (2015).
- Stern, S. A., Porter, S. B. & Zangari, A. M. On the roles of escape erosion and the viscous relaxation of craters on Pluto. *Icarus* **250**, 287–293 (2015).
- Eluszkiewicz, J. & Stevenson, D. J. Rheology of solid methane and nitrogen: application to Triton. *Geophys. Res. Lett.* **17**, 1753–1756 (1990).
- Protopapa, S. *et al.* Methane to nitrogen mixing ratio across the surface of Pluto. *Proc. Lunar Planet. Sci. Conf.* **47**, abstr. 2815 (2016).
- Eluszkiewicz, J. On the microphysical state of the surface of Triton. *J. Geophys. Res.* **96**, 19217–19229 (1991).
- Solomatov, V. S. Scaling of temperature- and stress-dependent viscosity convection. *Phys. Fluids* **7**, 266–274 (1995).
- Stansberry, J. A. & Yelle, R. V. Emissivity and the fate of Pluto's atmosphere. *Icarus* **141**, 299–306 (1999).
- Scott, T. A. Solid and liquid nitrogen. *Phys. Rep. (Phys. Lett. C)* **27**, 89–157 (1976).
- McKinnon, W. B., Simonelli, D. & Schubert, G. in *Pluto and Charon* (eds Stern, S. A. & Tholen, D. J.) 259–343 (Univ. Arizona Press, 1997).
- Robuchon, G. & Nimmo, F. Thermal evolution of Pluto and implications for surface tectonics and a subsurface ocean. *Icarus* **216**, 426–439 (2011).
- Lodders, K. Solar System abundances and condensation temperatures of the elements. *Astrophys. J.* **591**, 1220–1247 (2003).
- McKinnon, W. B. *et al.* The Pluto-Charon system revealed: geophysics, activity, and origins. *Lunar Planet. Sci. Conf.* **47**, abstr. 1995 (2016).
- Robuchon, G., Nimmo, F., Roberts, J. & Kirchhoff, M. Impact basin relaxation at Iapetus. *Icarus* **214**, 82–90 (2011).
- Estève, D. & Sullivan, N. S. NMR study of self-diffusion in solid N₂. *Solid State Commun.* **39**, 969–971 (1981).
- Moresi, L.-N. & Solomatov, V. S. Numerical investigation of 2D convection with extremely large viscosity variations. *Phys. Fluids* **7**, 2154–2162 (1995).
- Bland, M. T. & McKinnon, W. B. Forming Ganymede's grooves at smaller strain: toward a self-consistent local and global strain history for Ganymede. *Icarus* **245**, 247–262 (2015).
- Singer, K. N. & Stern, S. A. On the provenance of Pluto's nitrogen (N₂). *Astrophys. J.* **808**, L50 (2015).
- Gladstone, G. R. *et al.* The atmosphere of Pluto as observed by New Horizons. *Science* **351**, aad8866 (2016).
- Brown, M. E. in *The Solar System Beyond Neptune* (eds Barucci, M. A., Boehnhardt, H., Cruikshank, D. & Morbidelli, A.) 335–344 (Univ. Arizona Press, 2008).

Acknowledgements New Horizons was built and operated by the Johns Hopkins Applied Physics Laboratory (APL) in Laurel, Maryland, USA, for NASA. We thank the many engineers who have contributed to the success of the New Horizons mission, and NASA's Deep Space Network (DSN) for a decade of support of New Horizons. This work was supported by NASA's New Horizons project.

Author Contributions W.B.M. led the study and wrote the paper, with substantial input from F.N.; T.W. and J.H.R. performed the CitCom finite element convection calculations; P.M.S. developed the software to create stereographic and photoclinometric digital elevation models (DEMs) using New Horizons LORRI and MVIC images, and created the preliminary DEM for SP; O.L.W. mapped the SP region using New Horizons images in ArcGIS; J.M.M., J.R.S., A.D.H., O.M.U. and S.A.S. contributed to the understanding of the multiple roles N₂ ice plays in the geology of SP and its environs. S.A.S., H.A.W., C.B.O., L.A.Y. and K.E.S. are the lead scientists of the New Horizons project. The entire Geology, Geophysics, and Imaging Theme Team (listed) contributed to the success of the Pluto encounter.

Author Information All spacecraft data and higher-order products presented in this Letter will be delivered to NASA's Planetary Data System (<https://pds.nasa.gov>) in a series of stages in 2016 and 2017 because of the time required to fully downlink and calibrate the data set. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to W.B.M. (mckinnon@wustli.edu).

The New Horizons Geology, Geophysics and Imaging Theme Team

J. M. Moore (Lead)¹, W. B. McKinnon (Deputy Lead)², J. R. Spencer (Deputy Lead)³, R. Beyer¹, M. Buie³, B. Buratti⁴, A. Cheng⁵, D. Cruikshank¹, C. Dalle Ore¹, R. Gladstone⁶, W. Grundy⁷, A. D. Howard⁸, T. Lauer⁹, I. Linscott¹⁰, Francis Nimmo¹¹, C. Olkin³, J. Parker³, S. Porter³, H. Reitsema¹², D. Reuter¹³, J. H. Roberts⁵, S. Robbins³, P. M. Schenk¹⁴, M. Showalter¹⁵, K. Singer³, D. Strobel¹⁶, M. Summers¹⁷, L. Tyler¹⁰, H. Weaver⁵, O. L. White¹, O. M. Umurhan¹, M. Banks¹⁸, O. Barnouin¹⁹, V. Bray¹⁹, B. Carcich²⁰, A. Chaikin²¹, C. Chavez¹, C. Conrad³, D. Hamilton²², C. Howett³, J. Hofgartner²⁰, J. Kammer³, C. Lisse⁵, A. Marcotte⁵, A. Parker³, K. Retherford⁶, M. Saino⁵, K. Runyon⁴, E. Schindhelm³, J. Stansberry²³, A. Steffl³, T. Stryck²⁴, H. Throop³, C. Tsang³, A. Verbiscer⁸, H. Winters⁵, A. Zangari³, S. A. Stern³, H. A. Weaver⁵, C. B. Olkin³, L. A. Young³ & K. E. Smith¹

¹National Aeronautics and Space Administration (NASA) Ames Research Center, Moffett Field, California 94035, USA. ²Department of Earth and Planetary Sciences and McDonnell Center for the Space Sciences, Washington University in St Louis, Saint Louis, Missouri 63130, USA. ³Southwest Research Institute, Boulder, Colorado 80302, USA. ⁴NASA Jet Propulsion Laboratory, Pasadena, California 91019, USA. ⁵Johns Hopkins University Applied Physics Laboratory, Laurel, Maryland 20723, USA. ⁶Southwest Research Institute, San Antonio, Texas 78238, USA. ⁷Lowell Observatory, Flagstaff, Arizona 86001, USA. ⁸University of Virginia, Charlottesville, Virginia 22904, USA. ⁹National Optical Astronomy Observatory, Tucson, Arizona 85719, USA. ¹⁰Stanford University, Stanford, California 94305, USA. ¹¹Department of Earth and Planetary Sciences, University of California Santa Cruz, Santa Cruz, California 95064, USA. ¹²B612 Foundation, Mill Valley, California 94941, USA. ¹³NASA Goddard Space Flight Center, Greenbelt, Maryland 20771, USA. ¹⁴Lunar and Planetary Institute, Houston, Texas 77058, USA. ¹⁵The SETI Institute, Mountain View, California 94043, USA. ¹⁶The Johns Hopkins University, Baltimore, Maryland 21218, USA. ¹⁷George Mason University, Fairfax, Virginia 22030, USA. ¹⁸Planetary Science Institute, Tucson, Arizona 85719, USA. ¹⁹University of Arizona, Tucson, Arizona 85721, USA. ²⁰Cornell University, Ithaca, New York 14853, USA. ²¹Arlington, Vermont 05250, USA. ²²University of Maryland, College Park, Maryland 20742, USA. ²³Space Telescope Science Institute, Baltimore, Maryland 21218, USA. ²⁴Roane State Community College, Oak Ridge, Tennessee 37830, USA.

METHODS

Mapping and topography. The LORRI basemap in Fig. 1a was created from the 5×4 mosaic sequence P_LORRI (890 m per pixel), taken by the New Horizons Long-Range Reconnaissance Imager (LORRI). Mapping of cell/polygon boundaries (Fig. 1c) was carried out in ArcGIS using this mosaic and additional images from P_LORRI_Stereo_Mosaic (390 m per pixel). Figure 1a–c shows simple cylindrical projections, so the scale bars are approximate. Locations of Fig. 1d and Fig. 2a, b are shown as insets in Fig. 1a. Figure 2a is part of P_MVIC_LORRI_CA (MVIC Pan 2, 320 m per pixel), whereas Fig. 2b is a segment of the LORRI portion of P_MVIC_LORRI_CA, the highest resolution image transect obtained at Pluto by New Horizons (80 m per pixel).

Stereo topography over Sputnik Planum (SP) and its environs was determined using the two highest resolution Multispectral Visible Imaging Camera (MVIC) scans, P_MPan1 (495 m per pixel) and P_MVIC_LORRI_CA (MVIC Pan 2, 320 m per pixel). As MVIC is a scanning imager, each line must be individually registered carefully and pointing must be accurately known for stereo reconstruction. For Fig. 1b, Pluto was assumed to be a sphere of 1,187-km radius¹, and elevations were determined using an automated stereo photogrammetry method based on scene-recognition algorithms²⁷. Spatial resolutions are controlled by the lower resolution MVIC scan and, using this method, are further reduced by a factor of three to five. Vertical precisions can be calculated through standard stereo technique from $m r_p (\tan e_1 + \tan e_2)$, where m is the accuracy of pixel matching (0.2–0.3), r_p is pixel resolution, and e_1 and e_2 are the emission angles of the stereo image pair. For Fig. 1b the precision is about 230 m, well suited for determining elevations of Pluto's mountains and deeper craters as well as the rim-to-floor depth of the SP basin. It is not sufficient to determine planum cell/polygon elevations. In the planum centre, the dearth of sufficient frequency topography inhibits closure of the stereo algorithm, hence the noise in the centre of SP in Fig. 1b.

The subtle topography of the raised cells within SP was determined from a preliminary photoclinometric (shape from shading) analysis (for example, ref. 28), and is subject to further refinement of the photometric function for the bright cellular plains. Photoclinometry offers high-frequency topographic data at spatial scales of image resolution, but can be poorly controlled over longer wavelengths. Photoclinometry is sensitive to inherent albedo variations, but can be especially useful for investigating features with assumed symmetry, such as impact craters, which allows a measure of topographic control. The ovalar domes and bounding troughs of the bright cellular plains within SP are such symmetric features, and intrinsic albedo variations are muted in the absence of dark knobs or blocks, so photoclinometry is well-suited to determining elevations across individual cells within the bright cellular plains (Figs 1d and 2b).

Critical Rayleigh numbers for convection. Solid state viscosities η generally follow an Arrhenius law $\eta \approx \exp(E^*/RT)$ for any given rheological mechanism, where E^* is the activation energy for the deformation mechanism in question, R is the gas constant, and T is absolute temperature. For any given temperature and stress, one deformation mechanism generally dominates over another²⁹. Critical Rayleigh number values Ra_{cr} for convection, for a layer heated from below with fixed upper and lower boundary temperatures, depend on the deformation mechanism (through the power-law exponent in the stress strain-rate relation n) and the viscosity contrast $\Delta\eta$ across the layer due to the temperature difference ΔT . In what follows we adopt an exponential viscosity law based on a linear expansion of the Arrhenius law in E^*/RT (the Frank–Kamenetskii approximation) to take advantage of previous theoretical and numerical work^{13,22,30,31}. This is also an good approximation for the problem at hand because the temperature and viscosity contrast across a layer of volatile ices on Pluto is limited by the surface temperature of the ices on Pluto (37 K at the time of the New Horizons encounter)^{1,25} and the melting temperature of N_2 ice (63.15 K)¹².

For an exponential viscosity law, the driving (exponential) rheological temperature scale is $\Delta T_{th} \approx RT_i^2/E^*$, where T_i is a characteristic internal temperature of the convecting layer. The viscosity ratio across the layer due to temperature is then defined as $\Delta\eta = \exp(\theta) = \exp(\Delta T/\Delta T_{th})$. Ra_{cr} is then approximated, for large θ and in which $T_i \approx$ the basal temperature T_b , by¹³

$$Ra_{cr}(n, \theta) \approx Ra_{cr}(n) \left[\frac{\exp(1)\theta}{4(n+1)} \right]^{2(n+1)/n} \quad (2)$$

where $Ra_{cr}(n)$ is the critical Rayleigh number for non-Newtonian viscosity with no temperature dependence (2,038 for $n = 1$ and 310 for $n = 2.2$, based on numerical results for rigid upper and lower layer or sublayer boundaries^{32,33}). For large θ , convection occurs in the stagnant lid regime, in which convective motions are limited to a sublayer below a rigid surface. This is not the regime SP operates in, but serves as a limiting case. The transition from stagnant lid to sluggish lid convection, which does apply to SP, occurs at $\theta \approx 9$, or $\Delta\eta \approx 10^4$, for $n = 1$, and at $\theta \approx 13.8$, or $\Delta\eta \approx 10^6$, for $n = 2.2$ (ref. 13). The other convective regime limit

is that of small viscosity contrast ($\Delta\eta \rightarrow 1$). For SP, with a rigid lower boundary and a free-slip upper boundary, Ra_{cr} in this limit should be 1,101 (ref. 34) and ~ 200 (estimate) for $n = 1$ and 2.2, respectively. We then estimate $Ra_{cr}(n, \theta)$ for the sluggish lid regime, following refs 13 and 30, by linearly extrapolating in $\log\Delta\eta$ – $\log Ra_b$ space between the small viscosity contrast limit and the transition to stagnant lid convection:

$$Ra_{cr}(1, \theta) \approx 1,100 \exp(\theta/1.78) \quad (3a)$$

$$Ra_{cr}(2.2, \theta) \approx 200 \exp(\theta/3.87) \quad (3b)$$

The minimum or critical volatile ice layer thickness D_{cr} above which convection can occur and below which it cannot follows as³¹

$$D_{cr} \approx \left[\frac{Ra_{cr} \kappa^{1/n} \exp(E^*/nRT_i)}{3^{(n+1)/2n} A^{1/n} \rho g \alpha \Delta T} \right]^{n/(n+2)} \quad (4)$$

where κ , ρ , and α are, respectively, the thermal diffusivity, density, and volume thermal expansion coefficient of the ice, and A is the pre-exponential coefficient in the stress strain-rate relationship. For N_2 ice, this is either measured directly⁴ or estimated theoretically¹². The numerical factor in the denominator comes from the definition of viscosity and the conversion from laboratory geometry (A is measured in uniaxial compression) to the generalized flow law. For sluggish lid convection, we approximate T_i as $T_b - \Delta T/2$, which is a slight underestimate for the problem under discussion, but one that makes D_{cr} in equation (4) an upper bound on the minimum thickness for convection.

Equation (4) does not explicitly depend on ice grain size d . The power-law exponent reported for nitrogen ice deformation ($n \approx 2.2$)⁴ suggests a grain-size sensitive regime such as a grain boundary sliding, as opposed to a purely dislocation creep or climb mechanism (which would be grain-size independent)³⁵. Grain sizes in the nitrogen ice deformation experiments were not reported⁴, but it was noted that the grain sizes of similar experiments on methane ice were a few mm. This is a not atypical grain size for convecting upper mantle rock, or deep polar glacial ice on Earth, and is plausible for convecting water ice within icy satellites of the outer Solar System³⁶, so without further information we utilize the deformation experiment results for nitrogen⁴ as is. Notably, however, in order for N_2 ice to be identified spectroscopically at all on Pluto, very long optical path lengths are required (> 1 cm)³⁷, so the grain sizes of the convecting ice within SP may be much larger than a few millimetres. Because grain-size-sensitive rheologies typically have viscosities proportional to d^2 or d^3 , the presumed N_2 ice in SP may be much more viscous than in the reported experiments⁴. On the other hand, the presence of convective cells in SP implies that the viscosity is not arbitrarily large. Grain sizes in the annealed, convecting ice are probably determined by stress levels and the presence of contaminants (such as bits of water ice or tholins) and minor phases (such as CH_4 -rich ice)³⁶. Diffusion creep is also grain-size dependent, and in evaluating N_2 diffusion creep for comparison with Fig. 3 we adopt $d = 1$ mm as a nominal value, noting that for volume diffusion D_{cr} scales as $d^{2/3}$. The minimum thickness for convection by volume diffusion would plot off the graph in Fig. 3 to the upper right for $d = 1$ mm. Only if d were much smaller would D_{cr} for volume diffusion be comparable to that shown in Fig. 3.

Regarding the potential role of CO ice in SP, we note the near-perfect solid solution between solid N_2 and CO, and close similarities in density, melting temperature and electronic structure¹⁵. Hence, if the deeper ice in SP were actually dominantly CO, it would behave much the same as pure N_2 ice, with the proviso that an N_2 –CO ice solid solution under Pluto conditions would, for CO fractions greater than 10%, crystallize in the ordered α -phase, as opposed to the disordered β -phase of N_2 . We expect α -phase CO to be stiffer than its β -phase counterpart, based on the viscosity differences between ordered and disordered water ice phases³⁸. We stress, however, that the surface of SP, whatever its precise composition, is itself not in the α -phase, for if so the 2.16- μ m N_2 absorption feature would not be observed³⁷.

Regarding the potential role of CH_4 ice in SP, deformation experiments indicate similar behaviour to that of N_2 ice, but CH_4 ice appears to be about 25 times more viscous than N_2 ice (that is, A is ~ 25 times larger at the same T and differential stress)⁴, and with a similar power-law index n . The minimum or critical D_{cr} for convection within SP from equation (4) would then be about double that in Fig. 3 if SP were in fact filled with CH_4 ice, so the convection hypothesis is just as valid for CH_4 ice as for N_2 ice. The geological and compositional data point to an N_2 -dominated layer, however, as discussed in the main text.

Applying rheological data obtained in laboratory conditions to geological problems often requires extrapolation to different stress and strain conditions. For convection these conditions are lower stresses and strain rates. This is true whether one is modelling convection in the mantle of the Earth or another terrestrial

planet (with peridotite), in the icy satellites of the giant planets (with water ice), or in the present case of Sputnik Planum (with volatile ices such as N_2). The extrapolation is valid if the same stress mechanism or mechanisms dominate at the extrapolated conditions^{38,39}. The n values reported for laboratory deformation of N_2 ice and CH_4 ice⁴ are low enough (2.2 ± 0.2 and 1.8 ± 0.2 , respectively) that it seems implausible that some power-law, dislocation mechanism ($n \sim 3-5$) becomes dominant at lower stresses. Rather, the only likely transition would be, depending on T , to volume or grain-boundary diffusion ($n = 1$), which we already consider. Regardless, our understanding of N_2 and other volatile ice rheology could be greatly improved, especially any dependence on grain size.

Solid N_2 material parameters for Fig. 2 are as follows: $\kappa = 1.33 \times 10^{-7} \text{ m}^2 \text{ s}^{-1}$, $\alpha = 2 \times 10^{-3} \text{ K}^{-1}$, $E^* = 3.5 \text{ kJ mol}^{-1}$ ($n = 2.2$), $E^* = 8.6 \text{ kJ mol}^{-1}$ ($n = 1$), $A = 3.73 \times 10^{-12} \text{ Pa}^{-2.2} \text{ s}^{-1}$ ($n = 2.2$), $A = 1.52 \times 10^{-7} \times (d/1 \text{ mm})^{-2} \times (T/50 \text{ K})^{-1} \text{ Pa}^{-1} \text{ s}^{-1}$ ($n = 1$), $\rho = 1,000 - 2.14(T - 36 \text{ K}) \text{ kg m}^{-3}$, and for the heat flow calculations, conductivity $k = 0.2 \text{ W m}^{-1} \text{ K}^{-1}$ (refs 4, 15, 21). Pluto's surface gravity is 0.617 m s^{-2} (ref. 1).

Convection simulations. Numerical convection calculations were carried out with the well-benchmarked fluid dynamics finite element code CitCom²². CitCom solves the equations of thermal convection of an incompressible fluid in the Boussinesq approximation and at infinite Prandtl number. CitCom can solve the thermal convection equations using an Arrhenius viscosity or an exponential law (the Frank–Kamenetskii approximation). We used this latter approximation here, for both Newtonian (stress-independent) and non-Newtonian viscosities, to best compare our results with those in the literature^{5,13,22,30}.

We first simulated solid state convection with a Rayleigh number $Ra = 2 \times 10^4$ but with a non-temperature-dependent viscosity, in a very wide, rectangular 32×1 domain, with $2,048 \times 64$ elements, to allow natural selection of convection cell aspect ratios (widths of convective cells divided by layer depth). Temperatures at the top and bottom of the domain were fixed. Free slip was assumed at the surface, no slip at the base (the volatile ice layer is in contact with a rigid, water-ice basement), and periodic, free-slip boundary conditions along the sides of the domain. Velocities normal to domain edges in all cases were zero. Simulations were allowed to reach steady state. Calculations were carried out for Newtonian, isoviscous flow, and for non-Newtonian ($n = 2.2$) flow, both with the same Rayleigh number. In both cases the planforms were characteristic of their entire respective domains, and the aspect ratios for the convective cells for both simulations were close to 1, as expected from theory and previous results. (For example, the critical wavelength at $Ra = Ra_{cr}$ for a plane layer heated from below, with boundary conditions appropriate for convection within SP, is 2.34 times the layer depth³⁴.)

A suite of calculations was then carried at a variety of Ra_b and top-to-bottom viscosity ratios $\Delta\eta = \exp(\theta) = \exp(E^*\Delta T/RT_b^2)$, where Ra_b is defined as the basal Ra (that is, T in equation (1) of the main text = T_b). Rectangular 12×1 domains, with 768×64 elements, were used, with the same boundary conditions as above. A smaller number of calculations were also run with a free-slip lower boundary, for benchmarking with examples presented in ref. 5, and to simulate convection where the SP ice is at or near melting at its base. All runs in this suite were Newtonian, and while convective aspect ratios were not predictable from theory alone, they were expected to be much greater than 1 (ref. 5). In all cases simulations were allowed to reach steady state, or if time-dependent, to reach characteristic state behaviour.

Our present survey covers a range of Ra_b between 10^4 and 10^6 , and a range in $\Delta\eta$ between 150 and 3,000. This reflects our judgment that the convective regime

represented by the cells in SP ranges from the obviously convectively unstable to the subcritical (that is, stable) at the periphery of the basin (for example, Fig. 2b). The transition from cellular to non-cellular plains could reflect several things, including shallowing of the volatile ice layer, lower heat flow, and in the case of non-Newtonian flow, an insufficient initial temperature perturbation^{13,31,33}. The simplest explanation, however, for smaller cell sizes with distance from the centre of SP (Fig. 1c), and then a transition to level plains (no cells) towards the south (for example, Fig. 2b), is that the SP basin is shallower towards its margins, and particularly shallow towards its southern margin. This is consistent with the expected basin topography created by an oblique impact to the SSW⁴⁰. The less well defined cellular structure in the very centre of SP may, in contrast, reflect the deeper centre of the basin, implying a larger Ra for the N_2 ice layer there and more chaotic, time dependent convection.

Our numerical simulations are carried out in terms of dimensionless parameters, and do not presuppose any particular values for the depth of the SP volatile ice layer or Pluto's heat flow, and so on. They can be dimensionalized to determine if various measurable or estimable quantities are matched or are at least self-consistent. Depths and lengths scale as D , velocities as κ/D , stresses as $\eta_b \kappa/D^2$ (η_b is the basal viscosity), and heat flows as $k\Delta T/D$ (ref. 22). For example, for a given simulation, D can be scaled from surface cell size. Then different heat flows imply different ΔT . At fixed D and Ra_b , η_b , stresses, and dynamic topography all scale with ΔT .

Code availability. CitCom is freely available, in the version CitComS, released under a General Public License and downloadable from the Computational Infrastructure for Geodynamics (<http://geodynamics.org>).

27. Schenk, P. M., Wilson, R. R. & Davies, A. G. Shield volcano topography and the rheology of lava flows on Io. *Icarus* **169**, 98–110 (2004).
28. Schenk, P. M. Thickness constraints on the icy shells of the Galilean satellites from a comparison of crater shapes. *Nature* **417**, 419–421 (2002).
29. Frost, H. J. & Ashby, M. F. *Deformation-Mechanism Maps: The Plasticity and Creep of Metals and Ceramics* (Pergamon, 1982).
30. Solomatov, V. S. & Moresi, L.-N. Three regimes of mantle convection with non-Newtonian viscosity and stagnant lid convection on the terrestrial planets. *Geophys. Res. Lett.* **24**, 1907–1910 (1997).
31. Barr, A. C. & McKinnon, W. B. Can Enceladus' ice shell convect? *Geophys. Res. Lett.* **34**, L09202 (2007).
32. Stengel, K. C., Oliver, D. S. & Booker, J. R. Onset of convection in a variable viscosity fluid. *J. Fluid Mech.* **120**, 411–431 (1982).
33. Solomatov, V. S. & Barr, A. C. Onset of convection in fluids with strongly temperature-dependent, power-law viscosity: 2. Dependence on the initial perturbation. *Phys. Earth Planet. Inter.* **165**, 1–13 (2007).
34. Schubert, G., Turcotte, D. L. & Olson, P. *Mantle Convection in the Earth and Planets* (Cambridge Univ. Press, 2001).
35. Goldsby, D. L. & Kohlstedt, D. L. Superplastic deformation of ice: experimental observations. *J. Geophys. Res.* **106**, 11017–11030 (2001).
36. Barr, A. C. & McKinnon, W. B. Convection in ice I shells and mantles with self-consistent grain size. *J. Geophys. Res.* **112**, E02012 (2007).
37. Cruikshank, D. P. et al. in *Pluto and Charon* (eds Stern, S. A. & Tholen, D. J.) 221–267 (Univ. Arizona Press, 1997).
38. Durham, W. B., Prieto-Ballesteros, O., Goldsby, D. L. & Kargel, J. S. Rheological and thermal properties of icy materials. *Space Sci. Rev.* **153**, 273–298 (2010).
39. Karato, S. & Wu, P. Rheology of the upper mantle: a synthesis. *Science* **260**, 771–778 (1993).
40. Elbeshhausen, D., Wünnemann, K. & Collins, G. S. The transition from circular to elliptical impact craters. *J. Geophys. Res.* **118**, 2295–2309 (2013).

Attosecond nonlinear polarization and light–matter energy transfer in solids

A. Sommer^{1,*}, E. M. Bothschafter^{1,2,*†}, S. A. Sato³, C. Jakubeit¹, T. Latka¹, O. Razskazovskaya¹, H. Fattahi¹, M. Jobst¹, W. Schweinberger^{1,2}, V. Shirvanyan¹, V. S. Yakovlev^{1,4}, R. Kienberger⁵, K. Yabana^{3,6}, N. Karpowicz¹, M. Schultze^{1,2} & F. Krausz^{1,2}

Electric-field-induced charge separation (polarization) is the most fundamental manifestation of the interaction of light with matter and a phenomenon of great technological relevance. Nonlinear optical polarization^{1,2} produces coherent radiation in spectral ranges inaccessible by lasers and constitutes the key to ultimate-speed signal manipulation. Terahertz techniques^{3–8} have provided experimental access to this important observable up to frequencies of several terahertz^{9–13}. Here we demonstrate that attosecond metrology¹⁴ extends the resolution to petahertz frequencies of visible light. Attosecond polarization spectroscopy allows measurement of the response of the electronic system of silica to strong (more than one volt per ångström) few-cycle optical (about 750 nanometres) fields. Our proof-of-concept study provides time-resolved insight into the attosecond nonlinear polarization and the light–matter energy transfer dynamics behind the optical Kerr effect and multi-photon absorption. Timing the nonlinear polarization relative to the driving laser electric field with sub-30-attosecond accuracy yields direct quantitative access to both the reversible and irreversible energy exchange between visible–infrared light and electrons. Quantitative determination of dissipation within a signal manipulation cycle of only a few femtoseconds duration (by measurement and *ab initio* calculation) reveals the feasibility of dielectric optical switching at clock rates above 100 terahertz. The observed sub-femtosecond rise of energy transfer from the field to the material (for a peak electric field strength exceeding 2.5 volts per ångström) in turn indicates the viability of petahertz-bandwidth metrology with a solid-state device.

Matter responds to electromagnetic radiation by a displacement of its electrons with respect to the nuclei, turning its atomic constituents into dipole antennas. The overall strength of these dipoles per unit volume is characterized by the polarization vector, \mathbf{P} . Its dependence on the incident electric field, $\mathbf{E}(t)$, describes the macroscopic material response. Its nonlinear component, \mathbf{P}_{NL} , constitutes the basis for manipulating the electronic and optical properties with the electric field of light^{1,2}. The energy transferred from the electromagnetic field to the medium per unit volume can be expressed as:

$$W(t) = \int_{-\infty}^t \mathbf{E}(t') \cdot \frac{d}{dt'} \mathbf{P}_{\text{NL}}(t') dt' \quad (1)$$

Here we assume that the contribution of linear polarization to $W(t)$ is negligible. This is a prerequisite for ultrahigh-rate signal manipulation, which relies on low dissipation. In fact, it is this dissipation that has limited the clock rate in contemporary integrated digital electronics to several gigahertz¹⁵ for more than a decade^{16,17}.

A substantial increase of the electronic processing speed requires a new paradigm that is capable of greatly reducing the dissipation

per switching cycle. Recent experiments indicated a possible way of advancing contemporary microwave electronics to the frequency of visible light by manipulating the electronic and optical properties of wide-bandgap materials with strong visible light fields at photon energies much smaller than the bandgap of the material^{18,19}. However, the crucial question of how the energy density deposited irreversibly per switching cycle, $W_{\text{irreversible}}$, relates to the reversible energy exchange per unit volume, $W_{\text{reversible}}$, could not be answered. Pushing the frontiers of information processing to optical frequencies requires minimizing $W_{\text{irreversible}}$ while keeping $W_{\text{reversible}}$ high enough for reliable signal processing. Insight into field–matter energy exchange at optical frequencies requires access to $W(t)$ on a sub-femtosecond scale.

To this end, we propagated a strong, linearly polarized field $E(t)$ and its strongly attenuated replica $E_{\text{ref}}(t) = \beta E(t)$ through a thin sample of a transparent wide-bandgap material, in our case fused silica, of thickness ℓ . The attenuation factor β is sufficiently small to prevent any observable nonlinear material response to $E_{\text{ref}}(t)$. Both transmitted waveforms are recorded in a measurement sequence as outlined in Fig. 1, once attenuated after and once attenuated before the sample by the same attenuation factor β . We show in Supplementary Information section 1 how a difference between these transmitted waves, $\Delta E(t) = E(\ell, t) - \beta^{-1} E_{\text{ref}}(\ell, t)$, directly yields the nonlinear polarization $P_{\text{NL}}(t)$ induced by the strong field $E(t)$; see Supplementary equation (13).

In our experiments, we focus few-cycle near-infrared waveforms carried at a wavelength of $\lambda = 750$ nm into thin fused silica samples ($\ell = 10 \mu\text{m}$); for details of the experimental setup and procedures see Supplementary Information section 2. The focus of the transmitted waveform (that is, the interaction region) is imaged into an attosecond streak camera^{20,21}. Here the temporal evolution of the transmitted electric fields $E(z = \ell, t)$ and $E_{\text{ref}}(z = \ell, t)$ (henceforth referred to as $E(t)$ and $E_{\text{ref}}(t)$) is sampled with sub-250-as extreme-ultraviolet pulses. The peak intensities I_{peak} of the strong and attenuated waves have been set to $(1.3 \pm 0.1) \times 10^{14} \text{ W cm}^{-2}$ and $(6.7 \pm 0.3) \times 10^{12} \text{ W cm}^{-2}$, respectively. Figure 2a compares the transmitted fields and reveals the evolution of the nonlinear phase shift, $\Delta\varphi_{\text{NL}}(t)$, induced by the strong field. $\Delta\varphi_{\text{NL}}(t)$ increases towards the pulse peak, tapers off on its tail and finally vanishes; see insets to Fig. 2a. For $E_{\text{peak}} \approx 2.6 \pm 0.1 \text{ V Å}^{-1}$, the induced phase shift at the field maximum amounts to $\Delta\varphi_{\text{max}} = 0.7 \pm 0.1$ rad, which translates into a change of the refractive index by $\Delta n \approx (0.9 \pm 0.1) \times 10^{-2}$.

The field-induced phase shift evaluated at the pulse centre, $\Delta\varphi_{\text{peak}}$, is depicted in Fig. 2b and exhibits a linear scaling with the applied peak intensity. In contrast to previous research^{22–24}, our time-resolved study reveals the absence of saturation of the optical Kerr effect up to $E_{\text{peak}} \approx 2.7 \text{ V Å}^{-1}$, close to the threshold for dielectric breakdown for few-cycle laser pulses. The Kerr nonlinearity therefore appears to be

¹Max-Planck-Institut für Quantenoptik, Hans-Kopfermann-Strasse 1, 85748 Garching, Germany. ²Fakultät für Physik, Ludwig-Maximilians-Universität, Am Coulombwall 1, 85748 Garching, Germany. ³Graduate School of Pure and Applied Sciences, University of Tsukuba, Tsukuba 305-8571, Japan. ⁴Center for Nano-Optics and Department of Physics and Astronomy, Georgia State University, Atlanta, Georgia 30303, USA. ⁵Physik-Department, Technische Universität München, James-Frank-Strasse 1, 85748 Garching, Germany. ⁶Center for Computational Sciences, University of Tsukuba, Tsukuba 305-8577, Japan. [†]Present address: Paul Scherrer Institut, 5232 Villigen, Switzerland.

*These authors contributed equally to this work.

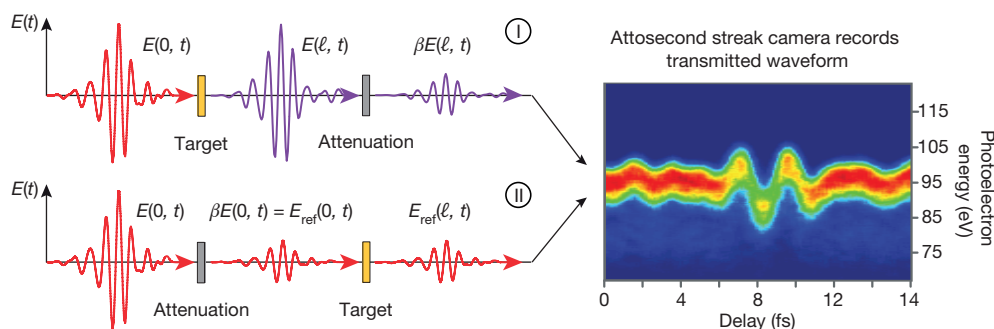


Figure 1 | Attosecond spectroscopy of the nonlinear polarization.

To induce a nonlinear material response $P_{NL}(t)$, the incident strong field $E(0, t)$ is transmitted through the sample, and subsequently its amplitude is decreased by the attenuation factor β before sampling the transmitted electric field waveform in a streak camera setup (I). The nonlinear polarization response is ‘deactivated’ by attenuating the

incident field before the sample and transmitting the weak reference field $E_{ref}(0, t) = \beta E(0, t)$ through the medium under scrutiny (II). The difference between the output waveforms, $\Delta E(t) = E(l, t) - \beta^{-1} E_{ref}(l, t)$, directly yields the nonlinear polarization of the medium, $P_{NL}(t)$, see Supplementary Information section 1. The false-colour plot shows a typical attosecond streaking spectrogram of the transmitted waveform used in the experiments.

potentially suitable for petahertz-scale signal manipulation and metrology beyond critical fields $E_{crit} \approx \Delta_g/(ea)$ (where Δ_g denotes the band-gap, $e = |e|$ is the elementary charge, and a is the lattice period; for silica, $E_{crit} \approx 2 \text{ V } \text{\AA}^{-1}$), provided that dissipation originating from carriers promoted into the conduction band during the nonlinear interaction

remains low. Although no lasting negative phase shift indicative of residual conduction band population is observable at the trailing edge of the waveform (where the Kerr effect vanishes), an accurate determination of the resultant $W_{irreversible}$ and the related $W_{reversible}$ requires evaluation of $P_{NL}(t)$.

The difference $\Delta E(t) = E(t) - \beta^{-1} E_{ref}(t)$ yields $P_{NL}(t)$ via Supplementary equation (13) (for details, see Supplementary Information section 1). Figure 3 depicts $P_{NL}(z = \ell/2, t)$ along with $E(z = \ell/2, t)$, both numerically propagated to the middle of the sample where their relative timing can be most precisely determined (see Supplementary Information section 1) for $E_{peak} = 2.6 \pm 0.1 \text{ V } \text{\AA}^{-1}$. $P_{NL}(t)$ oscillates almost perfectly in phase with $E(t)$, indicating a dominant role of bound electrons. This is in strong contrast to the response of free electrons appearing in the ionization of neon atoms in the gas phase²⁵, exhibiting a 90° phase shift with respect to the driving field (Supplementary Information section 3). A closer inspection reveals that $P_{NL}(t)$ lags slightly behind $E(t)$ on the front edge and the peak of the pulse, indicating—according to equation (1)—energy transfer from the field to the electronic system of fused silica, both of which become of opposite sign on the trailing edge of the pulse.

The response time of the polarizing electronic system, $\tau_{response}$, can be evaluated from the central zero-crossing of the fields (Fig. 3, upper left panel) as $\tau_{response} \approx 80$ as for $E_{peak} = 2.6 \pm 0.1 \text{ V } \text{\AA}^{-1}$. This is smaller than estimates from the Bohr orbit time¹ and from $\chi^{(3)}$ measurements in the range of 0.1–1 fs and decreases further with decreasing intensity, to well below 40 as for $E_{peak} < 2.2 \text{ V } \text{\AA}^{-1}$, as displayed in Fig. 3b. This can be understood by connecting $\tau_{response}$ to the nonlinear (field-induced) absorption coefficient, α_{NL} . For $\tau_{response}$ much smaller than the laser period, equation (1) yields a simple linear relationship, $\alpha_{NL} \propto \tau_{response}$. For multi-photon absorption, α_{NL} scales highly nonlinearly with the intensity and, according to this relationship, so does $\tau_{response}$. Supplementary Information section 4 presents detailed modelling of the intensity scaling of $\tau_{response}$ as well as a derivation of $\alpha_{NL}(\tau_{response})$.

Rendering $P_{NL}(t)$ an experimental observable, attosecond polarization spectroscopy allows to explore the intricate dynamic exchange of energy during nonlinear light–matter interactions. Inserting the measured values of $P_{NL}(t)$ and $E(t)$ into equation (1) provides direct experimental access to the work $W(t)$ done on the electrons by the laser field per unit volume, that is, the energy density transferred from the field to the electronic system. Figure 4a plots the measured $W(t)$ for several different peak intensities and the results of time-dependent density functional theory (TD-DFT) modelling (see inset to Fig. 4, ref. 26 and Supplementary Information section 5). We find very good qualitative agreement between theory and experiment regarding all observables analysed, including the behaviour of the maximum phase shift, the change in refractive index and the evaluated amount of dissipated energy. Quantitative agreement is achieved only when the

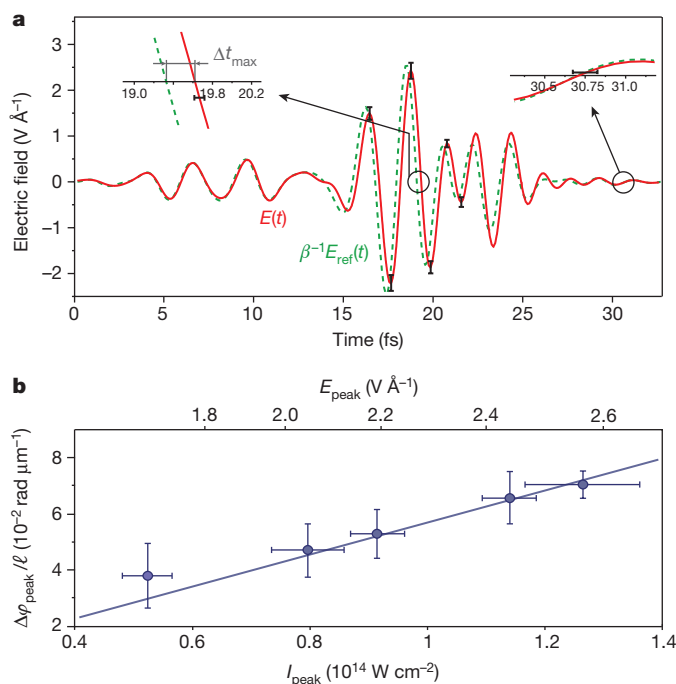


Figure 2 | Sub-femtosecond-resolved optical Kerr effect in silica.

a, After passage through a $10\text{-}\mu\text{m}$ -thick fused silica sample, the electric field $E(t)$ of the few-cycle near-infrared pulse with a peak intensity of $1.3 \times 10^{14} \text{ W cm}^{-2}$, approximately 10% below the threshold for optical damage, is modified as a result of the nonlinear light–matter interaction, as revealed by its comparison to a low-intensity ($I_{peak} = 7 \times 10^{12} \text{ W cm}^{-2}$) reference waveform $E_{ref}(t)$ (for $\beta = 0.27$). This comparison yields a transient positive phase shift induced by the strong field, as anticipated from the dynamic increase of the refractive index owing to the optical Kerr effect. The two insets show close-ups of the comparison near the centre and at the end of the pulse, revealing the full reversibility of the effect. $E(t)$ and $E_{ref}(t)$ are obtained from averaging a set of three recordings performed under identical conditions on individual samples. **b**, The phase shift $\Delta\varphi_{peak}$ evaluated at the peak of the field envelope for different peak intensities I_{peak} of $E(t)$ is found to exhibit a linear dependence on the field intensity. Each data point represents the mean value of three individual recordings under identical conditions; the error bars indicate the standard deviation.

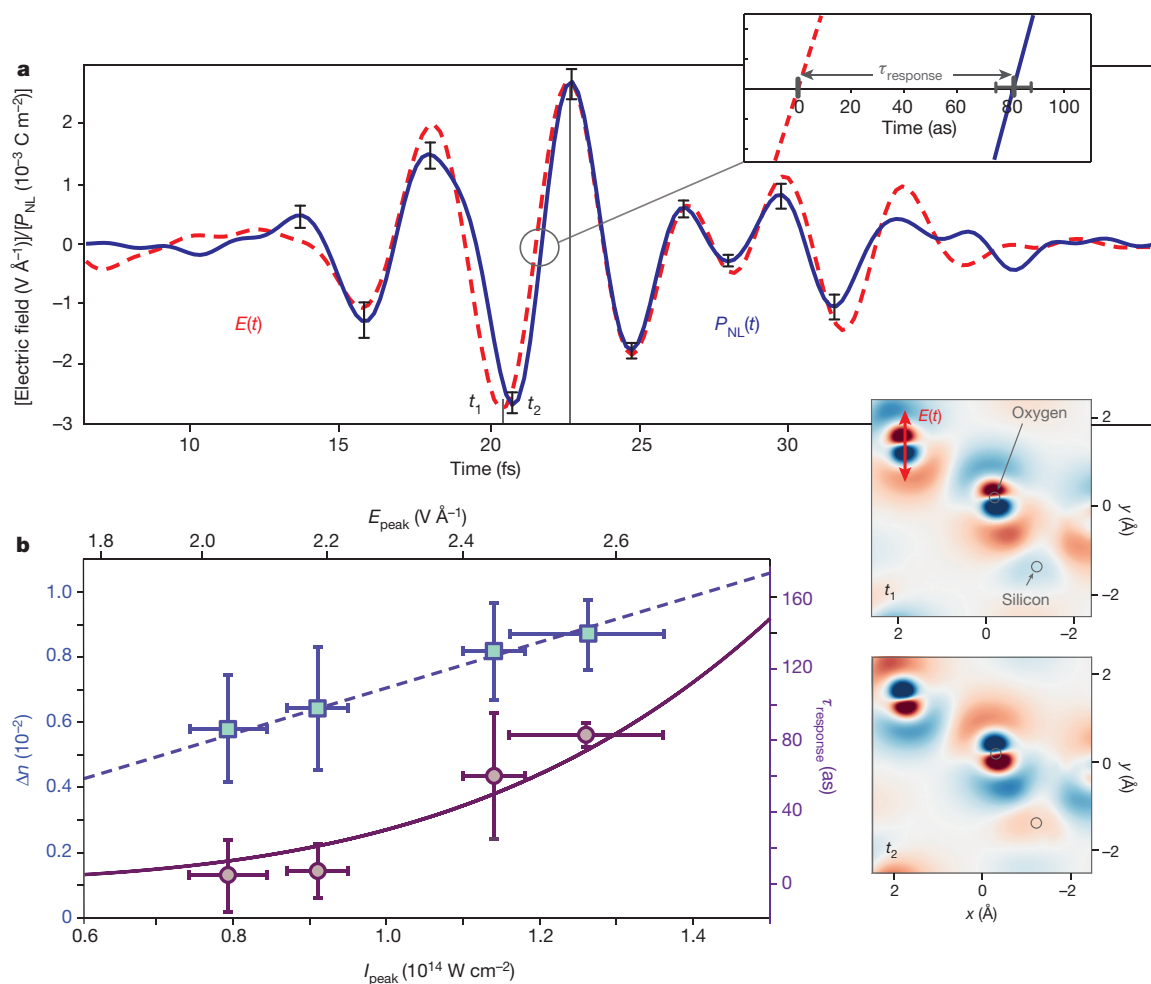


Figure 3 | The nonlinear optical polarization response of silica at critical field strengths. a, The strong electric field numerically back-propagated to the centre of the fused silica sample ($z = \ell/2 = 5 \mu\text{m}$) is contrasted with the nonlinear polarization $P_{\text{NL}}(t)$ evaluated from $E(t)$ and $E_{\text{ref}}(t)$ (Fig. 2a) at the same position (see Supplementary Information section 1). The response time of the nonlinear polarization near optical breakdown is found to be about 100 attoseconds at the pulse peak (close-up in the top inset). The other two insets display the computed spatial rearrangement of the electron density distribution for two extrema of the electric field at instants t_1 and t_2 in false-colour representation (red indicates an increase and blue a decrease relative to the unperturbed state). Electrons located in the vicinity of the oxygen atoms appear to dominate the polarization response, whereas the

electron cloud around the silicon centres remains largely unaffected. **b,** The response time of the nonlinear polarization is evaluated near the pulse peak as a function of the peak intensity of the applied field (circles) and compared to the results of a perturbation theory calculation (solid line; for details see Supplementary Information section 4). The field-induced change in refractive index Δn is evaluated from $P_{\text{NL}}(t)$ at the pulse peak as a function of the applied peak intensity I_{peak} . The nonlinear index n_2 determined from a linear regression (dashed line) is approximately one-third of the values acquired from time-integrated measurements using multi-cycle pulses²⁹. All data points and error bars represent the average and the standard deviation, respectively, of the evaluation of three individual data sets recorded under identical conditions.

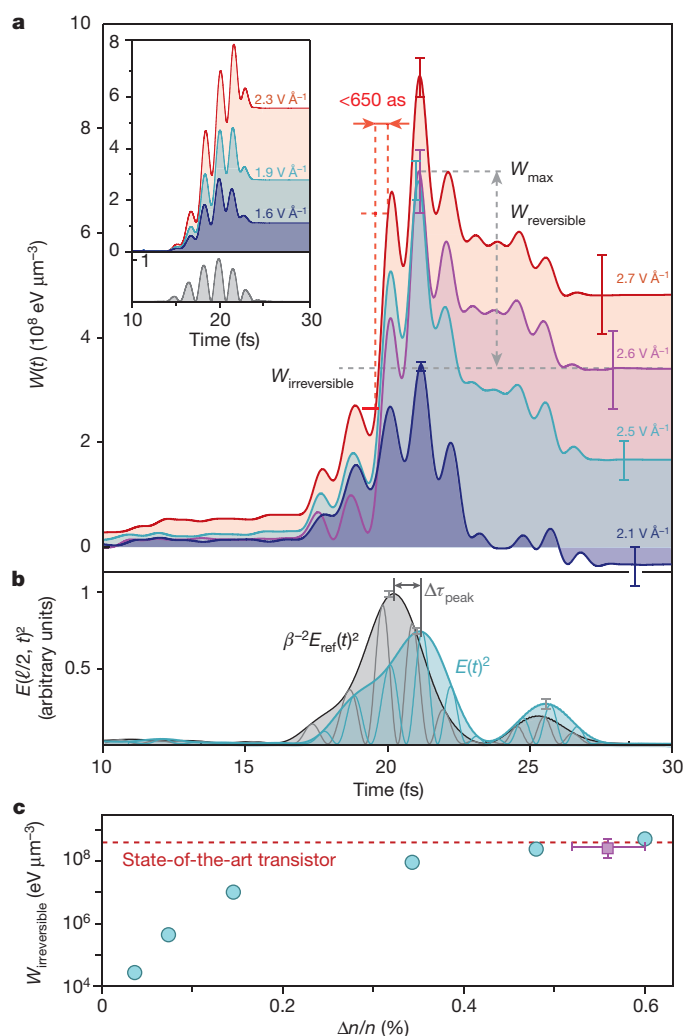
theoretically employed peak electric field is adjusted to values approximately 20% larger. This discrepancy can be attributed to inaccuracies of the exchange-correlation potential used in the TD-DFT calculations.

In all cases, the energy transferred from the field to the material increases up to the pulse peak and slightly beyond. This is because the field, while growing, needs to do ever more work to remove the electrons ever farther from their field-free location. The field amplitude decreasing after the pulse peak allows the displaced electrons to return gradually to their equilibrium position and radiate a part of the absorbed energy back into the driving laser field. This results in a negative slope for $W(t)$. The positive and negative slope are connected to the phase lag and phase advance of $P_{\text{NL}}(t)$ with respect to its driving field $E(t)$ before and after the pulse peak, respectively; these are clearly discernible in Fig. 3.

The energy density $W_{\text{irreversible}} = W(t \rightarrow \infty)$ irreversibly deposited in the system defines the charge carrier density promoted from the valence band into the conduction band according to $N_{\text{carrier}} \approx W_{\text{irreversible}}/\Delta_g$ (assuming population of the lowest-energy states of the conduction band). For $E_{\text{peak}} = 2.6 \text{ V } \text{\AA}^{-1}$ a residual

relative carrier concentration of $N_{\text{carrier}}/N_{\text{VB}} = 2.6 \times 10^{-4}$ is found, where $N_{\text{VB}} = 1.4 \times 10^{23} \text{ cm}^{-3}$ is the density of the valence-band electrons. This small residual carrier concentration is pivotal for future ultrafast signal processing and can hardly be determined with similar sensitivity by any other experimental method. By analogy, we can define the reversibly exchanged energy density as the difference between the maximum transferred energy, W_{max} , and $W_{\text{irreversible}}$. $W_{\text{reversible}} = W_{\text{max}} - W_{\text{irreversible}}$ can be interpreted in terms of a virtual conduction-band population with a number of virtual carriers of $N_{\text{virtual}} \approx W_{\text{reversible}}/\Delta_g$. N_{virtual} is the result of a projection of the laser-dressed and fully occupied valence-band states onto conduction-band states and hence it fully returns the energy density associated with it to the field upon its disappearance²⁷. In contrast, the real population, N_{carrier} , survives the field and—upon its subsequent decay—causes dissipation.

Although N_{virtual} seems rather elusive, attosecond metrology presents a very direct manifestation of the underlying reversible field–matter energy exchange. The initial energy flow dW/dt into the electronic system first extracts energy from the field on the leading edge of the



pulse. This is returned by a reversed energy flow on the trailing edge, resulting in a temporal shift of the pulse peak, $\Delta t_{\text{peak}} \approx 1 \text{ fs}$, as shown in Fig. 4b. The phenomenon is widely known as self-steepening or optical shock wave formation¹. Our study reveals that this phenomenon is an inherent consequence of the reversible field–matter energy transfer accompanying the field-induced change in the phase of the pulse. Hence, a field-induced change in the group index, Δn_g , is inextricably linked to that of the refractive index, Δn , implying a group delay and a phase shift, respectively.

Signal manipulation relies on the change of refractive index Δn (and Δn_g), which is characterized by $W_{\text{reversible}}$; in contrast, dissipation is detrimental to signal manipulation and is determined by $W_{\text{irreversible}}$. Hence, the scaling of $W_{\text{irreversible}}$ and $W_{\text{reversible}}$ (or, equivalently, Δn) with the applied field strength is of key importance for future signal-processing applications. We evaluated the dissipated energy per unit volume and per several-femtosecond optical switching/modulation cycle versus Δn from our *ab initio* TD-DFT calculations, which we verified against measurement at the highest field strength, near optical breakdown (see Fig. 4c and the discussion in Supplementary Information section 6). At measurable levels of Δn , $W_{\text{irreversible}}$ in a silica optical switch can be some four orders of magnitude smaller than the heat dissipation of a state-of-the-art metal oxide semiconductor field-effect transistor (MOSFET) operating at up to 10 GHz in integrated circuits. This very much reduced dissipation per switching cycle should therefore allow the operation of a dielectric switch/modulator at 100 THz or beyond.

An equally important discovery is the sub-femtosecond rise time of the transferred energy at $E_{\text{peak}} \geq 2.5 \text{ V } \text{\AA}^{-1}$ within each optical cycle.

Figure 4 | Energy exchange between strong optical fields and electrons in real time. **a**, The amount of energy the few-cycle near-infrared laser field transfers into a unit volume of silica is obtained from the measured $E(t)$ and $P_{\text{NL}}(t)$ via equation (1). $W(t)$ shows signatures of a substantial transient virtual conduction-band population (which is proportional to $W_{\text{max}} - W_{\text{irreversible}}$) oscillating in synchrony with the driving electric field. In the steepest of these oscillations, energy is transferred into the material within less than 650 as at $E_{\text{peak}} = 2.7 \text{ V } \text{\AA}^{-1}$. The amount of energy irreversibly dissipated in the sample $W_{\text{irreversible}}$ depends critically on the maximum applied field strength E_{peak} . Shown are the results of recordings for three different field amplitudes with $E_{\text{peak}} = 2.5 \text{ V } \text{\AA}^{-1}$, $2.6 \text{ V } \text{\AA}^{-1}$ and $2.7 \text{ V } \text{\AA}^{-1}$, as indicated, and a measurement closest to the average of five recordings with E_{peak} set equal to $2.1 \text{ V } \text{\AA}^{-1}$ (the uncertainty in the stated values of E_{peak} is $\pm 0.1 \text{ V } \text{\AA}^{-1}$). At this field strength, $W_{\text{irreversible}}$ becomes immeasurably small (with its error exceeding its nominal value). In the inset, $W(t)$ is computed from the nonlinear polarization at $z = \ell/2 = 5 \mu\text{m}$ obtained by the TD-DFT calculations outlined in Supplementary Information section 5 for a set of three different values of the peak electric field²⁶, as indicated. The spectrum of the computed nonlinear polarization shows the emergence of odd harmonics of the fundamental radiation (high harmonic generation). The results shown here are computed from the low-pass filtered nonlinear polarization to mimic the frequency transfer characteristics of the optical setup employed for the experiments. **b**, The squared electric field evolution of the reference wave and its envelope (black line) in comparison to the squared field and its envelope of the wave transmitted at a peak field strength of $2.5 \text{ V } \text{\AA}^{-1}$, showing clear indications of energy redistribution and consequent reshaping of the pulse envelope caused by the nonlinear polarization (see text). **c**, The dissipated energy density, equal to the irreversibly transferred energy density shown in **a**, as a function of the relative refractive index change, as extracted from the results of the TD-DFT simulation (circles) and experimental data taken at $E_{\text{peak}} = 2.6 \pm 0.1 \text{ V } \text{\AA}^{-1}$ (square). The excellent agreement between theory and experiment verifies the simulation results, permitting reliable prediction of the relevant quantities for much lower field strengths. The dashed line marks the dissipated energy density of a state-of-the-art MOSFET; for details see Supplementary Information Fig. 16. All error bars indicate the standard deviation of the evaluation of three sets of recordings performed under identical conditions.

With slightly shorter pulses than those used in these experiments²⁸, more than 90% of W_{max} will be transferred within a single sub-femtosecond rise depicted in Fig. 4. The resultant buildup of carriers in the conduction band within less than 1 fs will permit sampling of electric-field waveforms beyond the petahertz frontier in the simple setting demonstrated recently¹⁸.

Our proof-of-principle study on silica shows that careful choice of the peak electric field strength at $E < E_{\text{crit}}$ may open a route towards 100-THz-rate signal processing. The observed sub-femtosecond gradient in nonlinear energy transfer and the related change in electronic/optical properties at $E > E_{\text{crit}}$ may pave the way towards sampling optical fields (from the infrared to the ultraviolet) with a compact, cost-effective solid-state device. A petahertz solid-state oscilloscope should enable signal processing and metrology at visible light frequencies.

Traditional pump–probe spectroscopy makes use of the cycle-averaged amplitude envelope to resolve dynamics. In contrast, attosecond polarization spectroscopy uses the oscillating field as a probe, providing direct access to the full (linear and nonlinear) oscillating polarization and hence to the (reversible and irreversible) energy exchange between visible light and matter, as well as a delay in the system response. Hence, attosecond polarization spectroscopy is a generalization of pump–probe spectroscopy, yielding complete information about the dynamic electronic response of matter to strong visible light fields with attosecond resolution and, thanks to the intense attosecond field gradients, with a signal-to-noise ratio orders of magnitude better than that of any other attosecond technique demonstrated so far. Implemented with a probe waveform of sufficiently broad spectral

coverage, the approach allows, in principle, complete retrieval of the nonlinear polarization and hence of the entire response of the electronic system to strong-field excitation.

Received 16 August 2015; accepted 7 March 2016.

Published online 23 May 2016.

- Boyd, R. W. *Nonlinear Optics* (Academic Press, Elsevier, 2008).
- Wegener, M. *Extreme Nonlinear Optics: an Introduction* (Springer, 2005).
- Valdmanis, J. A., Mourou, G. & Gabel, C. W. Picosecond electro-optic sampling system. *Appl. Phys. Lett.* **41**, 211–212 (1982).
- Wu, Q. & Zhang, X. C. Free-space electro-optic sampling of terahertz beams. *Appl. Phys. Lett.* **67**, 3523 (1995).
- Sell, A., Leitenstorfer, A. & Huber, R. Phase-locked generation and field-resolved detection of widely tunable terahertz pulses with amplitudes exceeding 100 MV/cm. *Opt. Lett.* **33**, 2767–2769 (2008).
- Hebling, J., Lo Yeh, K., Hoffmann, M. C. & Nelson, K. A. High-power THz generation, THz nonlinear optics, and THz nonlinear spectroscopy. *IEEE J. Sel. Top. Quantum Electron.* **14**, 345–353 (2008).
- Huber, R. *et al.* Switching ultrastrong light–matter coupling on a subcycle scale. *J. Appl. Phys.* **109**, 102418 (2011).
- Leitenstorfer, A., Nelson, K. A., Reimann, K. & Tanaka, K. Focus on nonlinear terahertz studies. *New J. Phys.* **16**, 045016 (2014).
- Kuehn, W. *et al.* Terahertz-induced interband tunneling of electrons in GaAs. *Phys. Rev. B* **82**, 075204 (2010).
- Junginger, F. *et al.* Nonperturbative interband response of a bulk InSb semiconductor driven off resonantly by terahertz electromagnetic few-cycle pulses. *Phys. Rev. Lett.* **109**, 147403 (2012).
- Somma, C., Reimann, K., Flytzanis, C., Elsaesser, T. & Woerner, M. High-field terahertz bulk photovoltaic effect in lithium niobate. *Phys. Rev. Lett.* **112**, 146602 (2014).
- Ulbricht, R., Hendry, E., Shan, J., Heinz, T. F. & Bonn, M. Carrier dynamics in semiconductors studied with time-resolved terahertz spectroscopy. *Rev. Mod. Phys.* **83**, 543–586 (2011).
- Kampfrath, T., Tanaka, K. & Nelson, K. A. Resonant and nonresonant control over matter and light by intense terahertz transients. *Nature Photon.* **7**, 680–690 (2013).
- Krausz, F. & Ivanov, M. Attosecond physics. *Rev. Mod. Phys.* **81**, 163–234 (2009).
- Taur, Y. & Ning, T. H. *Fundamentals of Modern VLSI Devices* (Cambridge Univ. Press, 2009).
- Markov, I. L. Limits on fundamental limits to computation. *Nature* **512**, 147–154 (2014).
- Ionescu, A. M. & Riel, H. Tunnel field-effect transistors as energy-efficient electronic switches. *Nature* **479**, 329–337 (2011).
- Schiffrin, A. *et al.* Optical-field-induced current in dielectrics. *Nature* **493**, 70–74 (2012).
- Schultze, M. *et al.* Controlling dielectrics with the electric field of light. *Nature* **493**, 75–78 (2012).
- Itatani, J. *et al.* Attosecond streak camera. *Phys. Rev. Lett.* **88**, 173903 (2002).
- Kienberger, R. *et al.* Atomic transient recorder. *Nature* **427**, 817–821 (2004).
- Pati, A. P., Wahyutama, I. S. & Pfeiffer, A. N. Subcycle-resolved probe retardation in strong-field pumped dielectrics. *Nature Commun.* **6**, 7746 (2015).
- Loriot, V., Hertz, E., Faucher, O. & Lavorel, B. Measurement of high order Kerr refractive index of major air components. *Opt. Express* **17**, 13429–13434 (2009); erratum **18**, 3011–3012 (2010).
- Brée, C., Demircan, A. & Steinmeyer, G. Saturation of the all-optical Kerr effect. *Phys. Rev. Lett.* **106**, 183902 (2011).
- Geissler, M. *et al.* Light propagation in field-ionizing media: extreme nonlinear optics. *Phys. Rev. Lett.* **83**, 2930–2933 (1999).
- Yabana, K., Sugiyama, T., Shinohara, Y., Otobe, T. & Bertsch, G. Time-dependent density functional theory for strong electromagnetic fields in crystalline solids. *Phys. Rev. B* **85**, 045134 (2012).
- Yablonovitch, E., Heritage, J. P., Aspnes, D. E. & Yafet, Y. Virtual photoconductivity. *Phys. Rev. Lett.* **63**, 976–979 (1989).
- Cavaleri, A. L. *et al.* Intense 1.5-cycle near infrared laser waveforms and their use for the generation of ultra-broadband soft-x-ray harmonic continua. *New J. Phys.* **9**, 242 (2007).
- Milam, D. Review and assessment of measured values of the nonlinear refractive-index coefficient of fused silica. *Appl. Opt.* **37**, 546–550 (1998).

Supplementary Information is available in the online version of the paper.

Acknowledgements We acknowledge discussions with M. Stockman and V. Apalkov. This work was supported by the Max Planck Society and the Deutsche Forschungsgemeinschaft Cluster of Excellence: Munich Centre for Advanced Photonics (<http://www.munich-photonics.de>). M.S. was supported by a Marie Curie International Outgoing Fellowship (FP7-PEOPLE-2011-IOF). E.M.B. acknowledges funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement number 290605 (PSI-FELLOW/COFUND) and from the Swiss National Science Foundation through NCCR MUST. This research is based upon work supported by the US Air Force Office of Scientific Research under award number FA9550-16-1-0073 and used computational resources of the K computer provided by the RIKEN Advanced Institute for Computational Science through the HPCI System Research project (Project ID: hp140103).

Author Contributions F.K. and M.S. initiated, conceived and supervised the study. A.S. and E.M.B. developed the experimental method. A.S., E.M.B. and C.J. (in close cooperation with T.L., O.R., M.J., W.S. and V.S.) prepared and performed the experiment. S.A.S., H.F., K.Y. and N.K. accomplished the theoretical modelling. A.S., E.M.B., V.S.Y., R.K., N.K., M.S. and F.K. analysed and interpreted the experimental data. All authors discussed the results and contributed to the final manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to M.S. (martin.schultze@mpq.mpg.de) or F.K. (ferenc.krausz@mpq.mpg.de).

Oil sands operations as a large source of secondary organic aerosols

John Liggio¹, Shao-Meng Li¹, Katherine Hayden¹, Youssef M. Taha², Craig Stroud¹, Andrea Darlington¹, Brian D. Drollette³, Mark Gordon¹, Patrick Lee¹, Peter Liu¹, Amy Leithead¹, Samar G. Moussa¹, Danny Wang¹, Jason O'Brien¹, Richard L. Mittermeier¹, Jeffrey R. Brook¹, Gang Lu¹, Ralf M. Staebler¹, Yuemei Han¹, Travis W. Tokarek², Hans D. Osthoff², Paul A. Makar¹, Junhua Zhang¹, Desiree L. Plata³ & Drew R. Gentner³

Worldwide heavy oil and bitumen deposits amount to 9 trillion barrels of oil distributed in over 280 basins around the world¹, with Canada home to oil sands deposits of 1.7 trillion barrels². The global development of this resource and the increase in oil production from oil sands has caused environmental concerns over the presence of toxic compounds in nearby ecosystems^{3,4} and acid deposition^{5,6}. The contribution of oil sands exploration to secondary organic aerosol formation, an important component of atmospheric particulate matter that affects air quality and climate⁷, remains poorly understood. Here we use data from airborne measurements over the Canadian oil sands, laboratory experiments and a box-model study to provide a quantitative assessment of the magnitude of secondary organic aerosol production from oil sands emissions. We find that the evaporation and atmospheric oxidation of low-volatility organic vapours from the mined oil sands material is directly responsible for the majority of the observed secondary organic aerosol mass. The resultant production rates of 45–84 tonnes per day make the oil sands one of the largest sources of anthropogenic secondary organic aerosols in North America. Heavy oil and bitumen account for over ten per cent of global oil production today⁸, and this figure continues to grow⁹. Our findings suggest that the production of the more viscous crude oils could be a large source of secondary organic aerosols in many production and refining regions worldwide, and that such production should be considered when assessing the environmental impacts of current and planned bitumen and heavy oil extraction projects globally.

In general, secondary organic aerosol (SOA) mass is formed from the oxidation of organic gases, producing new compounds of sufficiently low saturation concentration (C^*) that can nucleate or condense onto pre-existing particles. SOA typically dominates total organic aerosol (OA) mass, and can account for >50% of particulate matter mass below 2.5 μm ($\text{PM}_{2.5}$) at many locations in the northern hemisphere¹⁰. SOA is partially derived from the oxidation of routinely measured volatile organic compounds (VOCs; $C^* > 10^6 \mu\text{g m}^{-3}$). However, recent evidence^{11,12} suggests that semi-volatility compounds (SVOCs; $C^* = 10^{-1} - 10^3 \mu\text{g m}^{-3}$) and intermediate-volatility compounds (IVOCs; $C^* = 10^3 - 10^6 \mu\text{g m}^{-3}$) are also important aerosol precursors owing to their high aerosol yields¹³. While oil and gas production and processing, including oil sands (OS) production, are known sources of VOC emissions¹⁴, their SVOC and IVOC emissions are unquantified. This is particularly relevant for the OS, since the mined material is a mixture of sand, water and clay coated in bitumen, the latter being an extremely viscous (and low-volatility) form of petroleum recovered through surface mining. During the Deepwater Horizon (DWH) oil spill, SVOCs and IVOCs were the predominant precursors of SOA formed downwind of the spill¹⁵. Heavy oils and bitumen are comprised of lower-volatility hydrocarbons than DWH crude¹⁶, such that their extraction and processing might be expected to release a

disproportionately large fraction of SVOCs and IVOCs into the atmosphere compared to lighter crude oil. On average, $5.04 \times 10^6 \text{ m}^3 \text{ month}^{-1}$ of bitumen was produced from OS surface mining operations in 2013 (ref. 17); should it be even slightly volatilized during production, there would be a strong potential for large amounts of SOA to be formed downwind of the region. This SOA formation potential from SVOC and IVOC emissions is demonstrated later.

Three aircraft measurement flights (F1, F2, F3) were conducted in Lagrangian patterns (Extended Data Fig. 1 and Supplementary Table 1), in which the same plume from OS operations was repeatedly sampled along tracks perpendicular to the plume axis (see Methods). Each flight intercepted two large, well-mixed plumes, revealing rapid SOA formation during transport, as illustrated in Extended Data Fig. 2 for F1 (similarly observed during F2 and F3). One plume was dominated by SO_2 and sulfate aerosols and the other by OA. While the sulfur plume can be traced back to OS facility stack emissions associated with desulfurization of raw bitumen, the origin of the large OA plume was less clear, and yet OA accounted for >80% of the aerosol mass (Extended Data Fig. 2). As the aircraft flew to different downwind distances from the OS (screens A, B, C and D), peak OA mass increased from ~ 10 to $14 \mu\text{g m}^{-3}$ (A to B) and remained constant at $\sim 12 \mu\text{g m}^{-3}$ (C to D), despite ongoing dilution (indicated by large decreases in SO_4^{2-} and black carbon (BC) aerosol concentrations), plume broadening (39 to 72 km) and particle deposition. This indicates a considerable SOA formation rate within these plumes, overriding the effect of dilution. Using BC as a tracer to correct for these effects (as described in Supplementary Discussion), a sixfold relative increase in OA mass (as SOA) is observed over 4 h (Fig. 1).

Net SOA formation rates were derived on the basis of mass balance using the OA mass transfer rates (tonnes (t) h^{-1}) across the flight screens¹⁸. The SOA formation rate is the OA transfer rate difference between screens. A description of the SOA production rate calculation, extrapolation assumptions and associated uncertainties is given in Methods. Accordingly, during F1, $3.4 \pm 0.9 \text{ t h}^{-1}$ of SOA was formed over $\sim 90 \text{ km}$ (A to D; Fig. 2), $2.7 \pm 1.0 \text{ t h}^{-1}$ between the screens of F2, and $2.1 \pm 0.9 \text{ t h}^{-1}$ during F3 (Extended Data Fig. 3). Including the SOA formed between the source region (S) and A, the cumulative SOA formation rates were 4.7 ± 0.9 , 5.3 ± 1.0 and $4.3 \pm 0.9 \text{ t h}^{-1}$ during F1, F2 and F3, respectively. Scaling by the time-integrated OH radical concentration over daylight hours, these formation rates translate to 45–84 t day^{-1} during the summer season. These remain underestimates since they do not include deposition or SOA formation beyond the last flight screens or at night. Correcting for depositional loss increases the rates to 55–101 t day^{-1} .

The rates of SOA formation observed here are very large; the relative rate of OA enhancement depicted in Fig. 1 is comparable to downwind of megacities such as Mexico City¹⁹ and Paris²⁰, and is higher than that

¹Air Quality Research Division, Environment and Climate Change Canada, Toronto, Ontario M3H 5T4, Canada. ²Department of Chemistry, University of Calgary, Calgary, Alberta T2N 1N4, Canada.

³Department of Chemical & Environmental Engineering, Yale University, New Haven, Connecticut 06520-8267, USA.

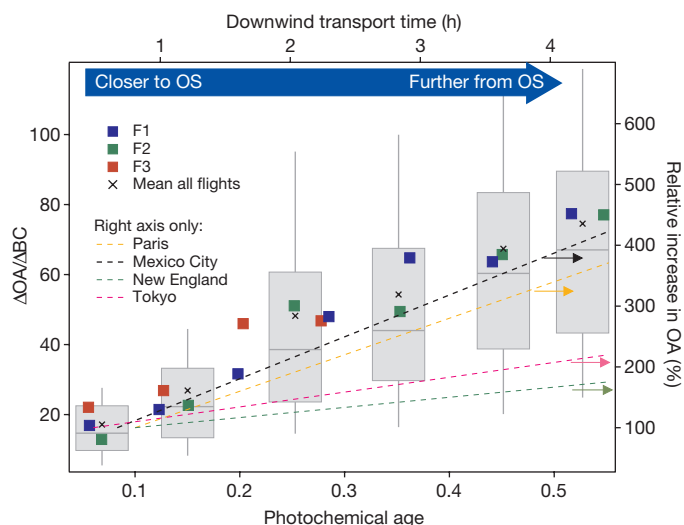


Figure 1 | Relative increase in OA downwind of the OS. The above-background (Δ)OA is normalized by BC (Δ OA/ Δ BC; left axis) and shown as a function of photochemical age ($-\log(\text{NO}_x/\text{NO}_y)$; bottom axis) and air mass transport time (top axis). Increases in Δ OA/ Δ BC indicate SOA formation. A sixfold relative increase in OA is observed (right axis), comparable to those reported downwind of large urban areas^{19–22}. Data points represent the average of the point-by-point Δ OA/ Δ BC binned by photo-chemical age. Grey boxes and whiskers represent 10th, 25th, 75th and 90th percentiles of the data from all three flights ($n = 2,573$).

observed in Tokyo²¹ and New England²², while the absolute rate (Fig. 2) is comparable to that estimated during the DWH oil spill ($\sim 3.3 \text{ t h}^{-1}$; ref. 15). However, a more compelling comparison to the absolute rate is with SOA formation rates downwind of major urban centres using available data (Fig. 2). For these urban centres, the SOA formed within one photochemical day was estimated using reported Δ OA/ Δ CO ratios and daily CO emissions, assuming that CO is co-emitted with SOA precursors^{23,24} (see Supplementary Discussion). The SOA formation rates downwind of the Greater Toronto Area (Canada's largest metropolis), Houston and the Mexico City Metropolitan area are estimated at 67, 52 and 228 t day^{-1} (not accounting for deposition), respectively. Despite the noted uncertainties described in Supplementary Discussion, this

comparison illustrates that OS operations are one of the largest sources of anthropogenic SOA in North America.

The SOA in these OS plumes had characteristics of two types of oxygenated organic aerosols (OOA)²⁵ as represented by two factors derived from positive matrix factorization (PMF) analysis of aerosol mass spectrometry data. Factor 1 (Extended Data Fig. 4) was more oxygenated than factor 2 (Fig. 3a), indicating that it was more photo-chemically aged. The time series of the factors during F1 are shown in Fig. 3b. Factor 1 was regionally distributed, dominating outside the plumes ($>80\%$) at $3\text{--}5 \mu\text{g m}^{-3}$, and largely consisted of aged regional biogenic SOA, as its mass spectrum was highly similar to those reported over forests²⁶ and from monoterpene oxidation in smog chamber experiments (Extended Data Fig. 4)²⁷. Factor 2 accounted for $>90\%$ of the SOA mass in the plume and was freshly formed from the oxidation of OS emissions. Its mass spectrum is almost identical to the spectra of OA derived from the OH oxidation of bitumen vapours in chamber experiments ($r^2 > 0.96$) (Fig. 3a and Extended Data Fig. 4), indicating that bitumen vapours are important precursors to the large SOA formation rates in OS plumes (see Supplementary Discussion).

The contribution of oxidized bitumen vapours to the observed SOA depends strongly on the initial volatility of the SOA precursors¹¹. To assess their SOA formation potentials, the volatility distributions (VDs) of bitumen vapours evolved from OS ore were determined (see Supplementary Methods), where the VD represents the fractions of total vapour in different ranges of C^* . At 20°C , the majority of vapour evolved is in the $C_{14}\text{--}C_{16}$ hydrocarbon range (IVOC; $C^* = 10^5 \mu\text{g m}^{-3}$), and shifts only slightly at 60°C (Fig. 4a). While gaseous emissions exist that span the $C_{12}\text{--}C_{18}$ range at ambient temperatures, heating of the material (70°C) results in complete evaporative loss up to C_{15} (Extended Data Fig. 5), leaving primarily compounds from C_{16} to $>C_{30}$. This represents a volatilization of $\leq 15\%$ of the total extractable hydrocarbon mass from the ore at 50°C , increasing further at higher temperatures (Fig. 4b). In surface mining operations, ore material is obtained via open-pit mining followed by bitumen-sand separation using hot water ($40\text{--}80^\circ\text{C}$) and further refining at up to 500°C . These derived bitumen vapour VDs clearly demonstrate the potential for atmospheric emissions of SOA precursors in a C^* range associated with strong SOA formation^{11,13}. On the basis of their volatility, such emissions are certain to occur during open-air mining and the various heated processing steps. Ambient ground-based measurements also show the existence of hydrocarbons

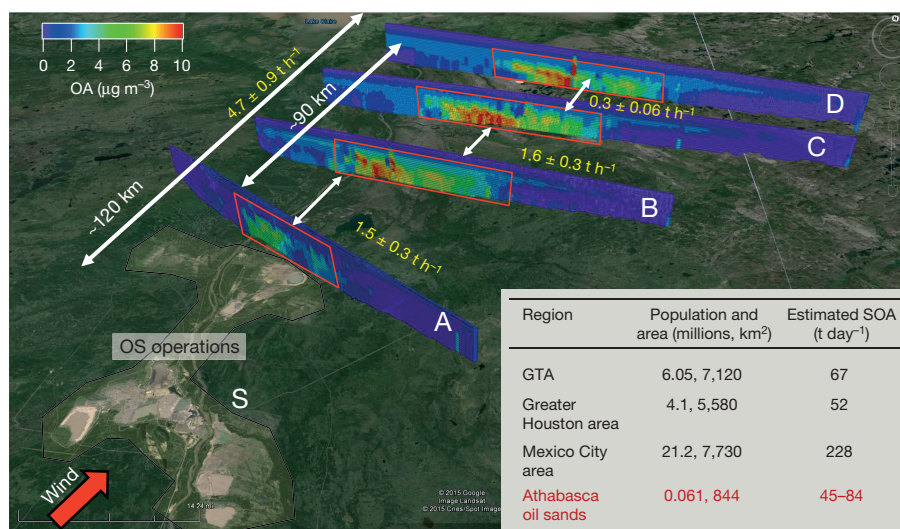


Figure 2 | OA mass screens during F1. SOA production is estimated as the sum of the differences in OA transfer rates between screens¹⁸. The overall rate from the source region (S) is the integrated OA transfer rate through screen D (4.7 t h^{-1}). SOA formed within ~ 1 photochemical day for major North American metropolitan areas is shown in the table,

compared to the range downwind of the OS (F1, F2, F3). Using Δ OA/ Δ CO to derive SOA for cities has been estimated to carry $\sim 50\%$ to $+100\%$ uncertainties²³. GTA, Greater Toronto Area. Map data: Google, image Landsat, Cnes/Spot Image 2015.

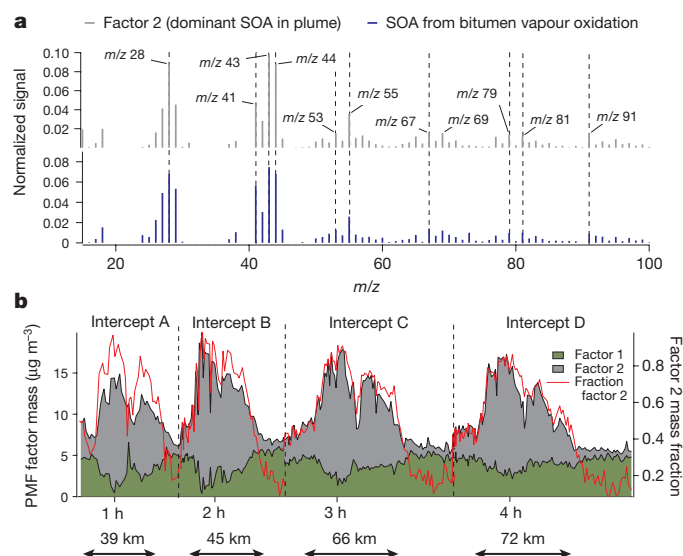


Figure 3 | PMF analysis for F1. **a**, PMF factor 2 profile during F1 compared to the mass spectra of SOA from the oxidation of bitumen vapours in a smog chamber, demonstrating a high degree of similarity ($r^2 = 0.96$). Signal is normalized to the total aerosol mass spectrometry (AMS) signal. **b**, Factor time series during F1 for consecutive plume intercepts approximately 1 h apart, at 600 m altitude. Factor 2 dominates the aerosol mass within the plume (red curve).

in this volatility range in plumes from OS facilities (Extended Data Fig. 6 and Supplementary Methods).

The bitumen SVOC and IVOC conversion to SOA in the observed plumes was further assessed with a Lagrangian box model constrained by the airborne measurements (Fig. 4c). The model simulated the formation of SOA in the plume of F1 over 3 h (screen A to D; Extended Data Fig. 2). Further details of the box model inputs and outputs are provided in Methods. From the ~ 70 p.p.b.v. of total VOCs measured at screen A, Fig. 4 demonstrates that only $< 6\%$ of the SOA after 3 h was contributed by the oxidation of speciated alkanes, alkenes and aromatic

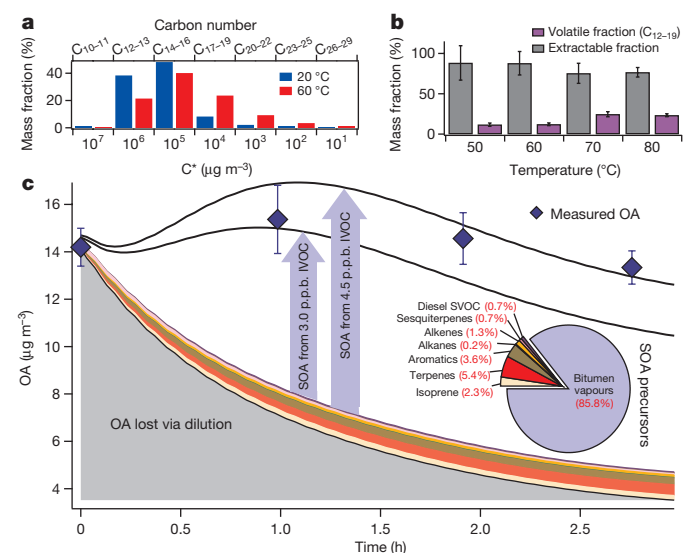


Figure 4 | Modelling SOA formation during F1. **a**, Volatility distribution of bitumen vapours at 20 °C and 60 °C. **b**, Fraction of the OS that is non-volatile (grey) and the volatile fraction (purple). Error bars represent standard deviation (s.d.) of $n = 3$ experiments. **c**, Box modelling of SOA formation during F1. A discrepancy between measured and modelled OA is reconciled by including 3.0–4.5 p.p.b.v. of bitumen IVOC vapours at time = 0 h (blue arrows). Error bars represent s.d. of the measured OA ($n = 7$). The pie chart indicates the contribution by each precursor type to the mass of SOA after 3 h.

hydrocarbons, and $< 9\%$ by isoprene and monoterpenes. The observed OA can only be reproduced by including bitumen SVOCs and IVOCs with the VD of Fig. 4a at 20 °C; adding 3–4.5 p.p.b.v. of bitumen SVOCs and IVOCs (with the current SOA ageing scheme used) at screen A adequately simulated the SOA measurements after 3 h (contributing $\sim 86\%$ of the SOA; Fig. 4c). Hence, even though the required SVOC and IVOC concentrations may be small (3–4.5 p.p.b.v.) compared to ~ 70 p.p.b.v. for VOCs, they dominate the contributions to SOA formation. Such a high SOA formation intensity is in contrast to most other types of energy production, which are likely to have emissions in a much lighter hydrocarbon range^{28,29}.

The evidence here indicates that large amounts of SOA will form from this previously unrecognized pool of OS-emitted SVOCs and IVOCs, dominating over SOA from traditional VOC precursors. The potential air-quality impacts of these vapours as a result of transport and refining could be more widespread than anticipated. Indeed, recent evidence indicates that primary IVOCs from an unknown petroleum-based source can account for about 30% of SOA mass in urban/suburban areas¹². This issue is not limited to Canada, as Venezuela plans to develop its Orinoco Oil Sands recoverable reserve of ~ 300 billion barrels, and the USA—having an estimated 54 billion barrel reserve of bitumen—has begun surface mining in Utah. In light of the current trend for increasing heavy oil production relative to conventional crude, further investigation is required to fully understand the magnitude of this potential global issue.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 9 November 2015; accepted 2 March 2016.

Published online 25 May 2016.

- Richard, F., Meyer, E. D. A. & Freeman, P. A. *Heavy Oil and Natural Bitumen Resources in Geological Basins of the World* (US Geological Survey, 2007).
- Government of Alberta. *Environmental Management of Alberta's Oil Sands* (Government of Alberta, 2009).
- Kelly, E. N. *et al.* Oil sands development contributes polycyclic aromatic compounds to the Athabasca River and its tributaries. *Proc. Natl Acad. Sci. USA* **106**, 22346–22351 (2009).
- Kirk, J. L. *et al.* Atmospheric deposition of mercury and methylmercury to landscapes and waterbodies of the Athabasca oil sands region. *Environ. Sci. Technol.* **48**, 7374–7383 (2014).
- Jung, K., Chang, S. X., Ok, Y. S. & Arshad, M. A. Critical loads and H⁺ budgets of forest soils affected by air pollution from oil sands mining in Alberta, Canada. *Atmos. Environ.* **69**, 56–64 (2013).
- Watmough, S. A., Whitfield, C. J. & Fenn, M. E. The importance of atmospheric base cation deposition for preventing soil acidification in the Athabasca Oil Sands Region of Canada. *Sci. Total Environ.* **493**, 1–11 (2014).
- Fuzzi, S. *et al.* Particulate matter, air quality and climate: lessons learned and future needs. *Atmos. Chem. Phys.* **15**, 8217–8299 (2015).
- BP. *Heavy Oil vs. Light Oil: Legislative Brown Bag* (BP, 2011).
- Dusseau, M. *Cold Heavy Oil Production with Sand in the Canadian Heavy Oil Industry* Ch. 2 (Alberta Energy, 2002).
- Zhang, Q. *et al.* Ubiquity and dominance of oxygenated species in organic aerosols in anthropogenically-influenced Northern Hemisphere midlatitudes. *Geophys. Res. Lett.* **34**, L13801 (2007).
- Gentner, D. R. *et al.* Elucidating secondary organic aerosol from diesel and gasoline vehicles through detailed characterization of organic carbon emissions. *Proc. Natl Acad. Sci. USA* **109**, 18318–18323 (2012).
- Zhao, Y. *et al.* Intermediate-volatility organic compounds: a large source of secondary organic aerosol. *Environ. Sci. Technol.* **48**, 13743–13750 (2014).
- Donahue, N. M., Robinson, A. L. & Pandis, S. N. Atmospheric organic particulate matter: from smoke to secondary organic aerosol. *Atmos. Environ.* **43**, 94–106 (2009).
- Simpson, I. J. *et al.* Characterization of trace gases measured over Alberta oil sands mining operations: 76 speciated C₂–C₁₀ volatile organic compounds (VOCs), CO₂, CH₄, CO, NO, NO₂, NO_y, O₃ and SO₂. *Atmos. Chem. Phys.* **10**, 11931–11954 (2010).
- de Gouw, J. A. *et al.* Organic aerosol formation downwind from the Deepwater Horizon oil spill. *Science* **331**, 1295–1299 (2011).
- Li, R. *et al.* Laboratory studies on secondary organic aerosol formation from crude oil vapors. *Environ. Sci. Technol.* **47**, 12566–12574 (2013).
- Alberta Energy Regulator. *Alberta Mineable Oil Sands Plant Statistics* (Alberta Energy Regulator, 2013).
- Gordon, M. *et al.* Determining air pollutant emission rates based on mass balance using airborne measurement data over the Alberta oil sands operations. *Atmos. Meas. Tech.* **8**, 3745–3765 (2015).

19. Kleinman, L. I. *et al.* The time evolution of aerosol composition over the Mexico City plateau. *Atmos. Chem. Phys.* **8**, 1559–1575 (2008).
20. Freney, E. J. *et al.* Characterizing the impact of urban emissions on regional aerosol particles: Airborne measurements during the MEGAPOLI experiment. *Atmos. Chem. Phys.* **14**, 1397–1412 (2014).
21. Miyakawa, T., Takegawa, N. & Kondo, Y. Photochemical evolution of submicron aerosol chemical composition in the Tokyo megacity region in summer. *J. Geophys. Res.* **113**, D14304 (2008).
22. Kleinman, L. I. *et al.* Aircraft observations of aerosol composition and ageing in New England and Mid-Atlantic States during the summer 2002 New England Air Quality Study field campaign. *J. Geophys. Res.* **112**, D09310 (2007).
23. de Gouw, J. & Jimenez, J. L. Organic aerosols in the Earth's atmosphere. *Environ. Sci. Technol.* **43**, 7614–7618 (2009).
24. Hayes, P. L. *et al.* Modeling the formation and aging of secondary organic aerosols in Los Angeles during CalNex 2010. *Atmos. Chem. Phys.* **15**, 5773–5801 (2015).
25. Ng, N. L. *et al.* Real-time methods for estimating organic component mass concentrations from aerosol mass spectrometer data. *Environ. Sci. Technol.* **45**, 910–916 (2011).
26. Robinson, N. H. *et al.* Evidence for a significant proportion of secondary organic aerosol from isoprene above a maritime tropical forest. *Atmos. Chem. Phys.* **11**, 1039–1050 (2011).
27. Liu, Y., Liggitto, J., Staebler, R. & Li, S. M. Reactive uptake of ammonia to secondary organic aerosols: kinetics of organonitrogen formation. *Atmos. Chem. Phys.* **15**, 13569–13584 (2015).
28. Gentner, D. R. *et al.* Emissions of organic carbon and methane from petroleum and dairy operations in California's San Joaquin Valley. *Atmos. Chem. Phys.* **14**, 4955–4978 (2014).
29. Gilman, J. B., Lerner, B. M., Kuster, W. C. & de Gouw, J. A. Source signature of volatile organic compounds from oil and natural gas operations in northeastern Colorado. *Environ. Sci. Technol.* **47**, 1297–1305 (2013).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank the National Research Council of Canada flight crew of the Convair-580, the technical support staff of the Air Quality Research Division, S. Cober for the management of the study, and the community of Fort McKay for the support of the Oski ôtin ground site at Fort McKay. The project was supported by the Clean Air Regulatory Agenda and the Joint Oil Sands Monitoring program.

Author Contributions All authors contributed to the collection of observations in the field, in the laboratory or the development of the box model. J.L. and S.-M.L. wrote the paper with input from all co-authors. S.-M.L. designed and directed the flights. Y.M.T. and C.S. conducted the box modelling work with input from J.L. D.R.G., D.P., B.D.D. and P.L. provided bitumen volatility distributions.

Author Information The data used are available on the Canada-Alberta Oil Sands Environmental Monitoring Information Portal (<http://jointoilsandsmonitoring.ca/default.asp?n=5F73C7C9-1&lang=en>). Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to J.L. (John.Liggitto@canada.ca) or S.-M.L. (Shao-Meng.Li@canada.ca).

METHODS

Aircraft campaign. Airborne measurements of an extensive set of air pollutants over the Athabasca oil sands region in northern Alberta were conducted between 13 August and 7 September 2013 in support of the Joint Canada-Alberta Implementation Plan on Oil Sands Monitoring. Instrumentation was installed aboard the National Research Council of Canada Institute for Aerospace Research (NRC Aerospace) Convair-580 research aircraft. The aircraft flew 22 flights over the Athabasca oil sands, for a total of approximately 84 h. Thirteen flights were designed specifically to quantify area emissions from various OS facilities by flying in a rectangular box shape, at multiple altitudes, resulting in 21 box flights around 7 different OS facilities.

A further three flights (denoted F1 (4 September), F2 (5 September) and F3 (19 August)) were designed to study the transformation of OS emitted pollutants, including the formation of SOA. These flights were designed as Lagrangian experiments in which the same air parcels in OS plumes were sampled at different time intervals (1 h apart) as the air parcels were transported downwind for 4–5 h. The measurement locations for the flight tracks were chosen so that the aircraft would intercept the same air parcel, using real-time wind speed/direction measurements to guide the intercept locations. The intercepting flight tracks were perpendicular to the axis of the plumes, and the flight times crossing the plumes were 5–7 min. At each intercept location, high time resolution (1 s for gases, 10 s for AMS measurements) measurements were made at multiple altitudes (2–5 horizontal transects) from ~150 m above ground to over 1,400 m, which was higher than the mixed layer height, consisting of level flight tracks and spirals at the centre of the plume. These vertically spaced level flight tracks and spirals constituted virtual screens at the intercept locations. The three flights (F1, F2 and F3) comprised 5, 3 and 3 screens, respectively. In between the screens in each flight, there were no industrial emissions. Thus, changes between screens can be described in terms of mixing/dilution, chemistry and deposition that occurred from within a single air parcel.

The first screens of the F1, F2 and F3 flights were approximately 1 h downwind of the majority of OS facilities, and at distances that pollutants from multiple OS sources were well mixed and merged into large plumes. The flight paths and their associated parameters are given in Extended Data Fig. 1 and Supplementary Table 1. As shown in this figure, the Lagrangian experiments resulted in varying degrees of success for a number of reasons, including data capture rates, consistency of winds, and the exact timing of when the aircraft crossed the plumes at the chosen intercepting locations, with F1 having the best matches between the air parcel transport times and the aircraft flight times at the screen locations. As a result, the data from F1 are used more extensively than others here, although not exclusively.

The Convair-580 was equipped with fast response instrumentation to measure an extensive set of gas- and particle-phase pollutants, as well as standard meteorological and aircraft state parameters. A description of the meteorological variables and aircraft state parameters measured is given elsewhere¹⁸. Non-refractory (NR) particle composition (that is, ammonium, nitrate, sulfate and organics) was measured with an Aerodyne high-resolution time-of-flight aerosol mass spectrometer (HR-ToF-AMS; Aerodyne Research)³⁰. Refractory black carbon (BC) particle measurements were made with a Single Particle Soot Photometer (SP2; Droplet Measurement Technologies)^{31,32}. A subset of volatile organic compounds (VOCs) was measured with a high-resolution proton transfer time-of-flight mass spectrometer (PTR-ToF-MS; Ionicon Analytik GmbH)³³ and a more extensive set of hydrocarbons was measured via on-board canister sampling, followed by analysis by gas chromatography mass spectrometry and flame ionization detection (GC-MS and GC-FID). A full description of all the relevant gas- and particle-phase instrumentation aboard the aircraft is provided in the Supplementary Information. No statistical methods were used to predetermine sample size.

OA mass transfer rate and OS SOA production rate calculations. The quantification of the mass transfer rate of organic aerosols (R_{OA} , in t h^{-1}) across a virtual screen uses an extension of the top-down emission rate retrieval algorithm (TERRA) described previously¹⁸. TERRA was originally developed to determine emission rates from box flight patterns during this study¹⁸, based on mass balance within the virtual box constructed from the flight tracks. Briefly, TERRA uses the flight path around a facility at multiple altitudes to map the data to the two-dimensional virtual walls of a box surrounding the facility. The transport of a pollutant through the walls is calculated using aircraft wind and compound mixing ratio measurements, and emission rates calculated on the basis of the divergence theorem with estimations of box-top loss rates, horizontal and vertical advective and turbulent transport rates, surface deposition rate, and apparent loss rates due to air densification and chemical reaction rates. For the transformation flights, some components of TERRA were extended to apply to single screens created from vertically stacked level flight tracks and spirals. Concentration data C (in $\mu\text{g m}^{-3}$) are mapped to the screens and interpolated using a simple kriging function (on approximately 5,000–15,000 individual data points). Wind speed along the flight tracks was decomposed into two components based on the wind direction,

one parallel to the screen (u_p) and the other normal to the screen (u_n), and the decomposed wind speeds were similarly mapped to the screen and interpolated using kriging. The lowest flight altitude was at approximately 150 m, hence there was a need to extrapolate the OA measurements and the wind speed components downward to the ground surface. The downward extrapolation for the wind speed components assumed a stability-dependent log profile³⁴ vertically and uses nearby concurrent wind profiler data to determine the roughness and displacement height¹⁸. The OA measurement downward extrapolation was based on the assumption of a well-mixed layer below the lowest flight track altitude, which is consistent with modelling³⁵ and the potential temperature profile. A variation to this downward extrapolation method assumed a linear downward trend from the flight altitudes, to capture possible variations in the mixing state below the lowest flight track altitude. Previous analysis has shown that unknown pollutant concentrations below the lowest flight level (and the associated extrapolation to ground) led to the majority of the uncertainty in the emissions estimates from this approach (~20%; ref. 18). The OA measurements during the flights here were extrapolated downward using both methods; varying linearly to the ground or held constant (at the lowest altitude concentration) to the ground, to assess the uncertainty in the final derived mass transfer rate caused by the extrapolation methods. The OA data were further linearly extrapolated from the highest altitude level flight tracks upwards (to background OA concentrations) in the case where the level flight tracks did not traverse vertically beyond the mixed layer. The highest altitude extrapolated to was determined from the OA measurements and temperature profiles from spirals along the tracks, which were flown above the top of the boundary layer but not included in the screens. The results showed a difference of <15% for the mass transfer rates among the different extrapolation schemes.

The mass transfer rate of OA across each screen (R_{OA}) of flights F1, F2 and F3 was derived on the basis of the extended TERRA as described earlier and the HR-ToF-AMS data. To avoid the background OA affecting the computation of R_{OA} , a background OA (Extended Data Fig. 7) was subtracted from the OA measurements in the following computation:

$$R_{OA}(A) = \int_{s_1}^{s_2} \int_{z_1}^{z_2} C(s, z, A) u_n(s, z, A) ds dz \quad (1)$$

where s_1 and s_2 are the horizontal edge positions on the screen for the plume containing OA, z_1 is the ground surface altitude, z_2 is the top of the plume, $C(s, z, A)$ is the interpolated/extrapolated concentration on screen A (and other screens), and $u_n(s, z, A)$ is the interpolated/extrapolated wind speed vector normal to screen A. The plume edges are determined by the OA concentration on the screen, indicated by $C(s, z, A)$, approaching the background concentration of approximately $4 \mu\text{g m}^{-3}$. Note that equation (1) describes horizontal advective transfer rates only; additional contribution from horizontal turbulent fluxes can contribute to R_{OA} but this has been shown to be a few orders of magnitude smaller than the horizontal advective transfer¹⁸ and therefore is ignored henceforth.

Between screens, the mass transfer rate R_{OA} may change due to emissions with a rate of E_{OA} , deposition with a rate of D_{OA} , and the formation of SOA at a rate of R_{SOA} . In the original TERRA, vertical advective and turbulent transfer rates as well as air density changes were considered to achieve mass balance when the background level of a compound was large¹⁸. The vertical transport term was nominally small compared to the horizontal advection, and hence can be ignored. Thus, using a mass balance approach, the following relationship can be established

$$R_{OA}(t_2) = R_{OA}(t_1) + R_{SOA} + E_{OA} - D_{OA} \quad (2)$$

where t_1 and t_2 are the times of the two screens where the plume parcels were intercepted. Positive matrix factorization (PMF) analysis of the HR-ToF-AMS data from the transformation flights F1, F2 and F3 showed no hydrocarbon-like aerosol factor²⁵, suggesting small-to-non-existent contributions from primary emissions of organic aerosols between the screens or from the source region to the screens. Hence $E_{OA} = 0$. Using concurrent refractory BC measured by SP2, the maximum dry deposition of BC over the region was estimated to be approximately $7\% \text{ h}^{-1}$ derived from the differences in the BC mass transfer rates across the screens. We assume that this rate of deposition of BC is applicable to OA. Since deposition derived this way is relatively small, it is ignored to derive the SOA formation rate according to

$$R_{SOA} \approx R_{OA}(t_2) - R_{OA}(t_1) \quad (3)$$

Equation (3) was used to calculate the SOA formation rates, ignoring the dry deposition term, to be comparable to urban SOA estimates, which are net of deposition. Including a fully evaluated dry deposition for the R_{SOA} calculation would mean that equation (3) gives a lower limit of the true SOA formation rate during the measurement period. The total SOA production rate (R_{SOA}) in these flights is taken to be the

OA transfer rate (R_{OA}) through the final screen, since $E_{OA} = 0$ and only oxygenated PMF factors were observed. The total SOA is then extrapolated to a photo-chemical day as described in Supplementary Discussion (Extended Data Fig. 8).

Box modelling description. SOA formation in the large-scale plume of F1 was modelled with a zero-dimensional Lagrangian box model, as it evolved over approximately 3 h (~600 m altitude). The simulation was constrained by the measurements of VOCs, NO_x , OVOCs, O_3 and other parameters, while dilution within the plume was accounted for using BC as a dilution tracer. Hydrocarbons of both anthropogenic and biogenic origin were constrained at the first screen (A), or throughout the simulation for those biogenic species with potential continuous emissions along the flight track (monoterpenes and isoprene). Background concentrations were constrained by measurements outside of the plume. The model uses the Statewide Air Pollution Research Centre (SAPRC07) chemical mechanism with updated isoprene chemistry^{36–38}. The model was run with a 2 min time step and diluted chemical species at every time step. While the model had VOCs constrained, including a constraint for NO_x and O_3 resulted in very little difference between the model and observations. Hence, the gas-phase chemistry is well simulated by the box model, as shown in Extended Data Fig. 9. Sesquiterpenes were constrained based on the ratio to measured monoterpenes. Sesquiterpenes were estimated from the PTR-ToF-MS measurements using an estimated ion transmission efficiency and proton transfer reaction kinetics, in a manner described previously^{39,40}, resulting in a sesquiterpene:monoterpene ratio of ~0.39. This is somewhat higher than the ratios of 0.013 and 0.105 that have been recommended previously^{41,42}, and was used as an upper estimate to the sesquiterpene contribution to SOA. Regardless, biogenic VOCs contributed little to the observed and modelled SOA (Extended Data Fig. 10 and Supplementary Discussion). Recent evidence has also suggested that extremely low-volatility compounds (ELVOC) can also form via an auto-oxidation mechanism⁴³. This process has been demonstrated to be most relevant in rural and remote regions where OA loading, VOC and NO_x levels are very low, due to competing $RO_2 + NO$ and/or $RO_2 + RO_2$ reactions. Previous data⁴³ indicate that ELVOC yields are most important at 1 p.p.b.v. NO_x and below. While ELVOC may be an important SOA contributor outside of the OS plumes (where biogenics are abundant and NO_x is low), the amount of NO_x in the OS plumes studied (as well as the OA loading and VOC levels) were far too high (approaching >20 p.p.b.v. NO_x and always greater than 1 p.p.b.v.) for ELVOC formation to be important. Hence, the contribution of ELVOC was not explicitly included in the box model analysis.

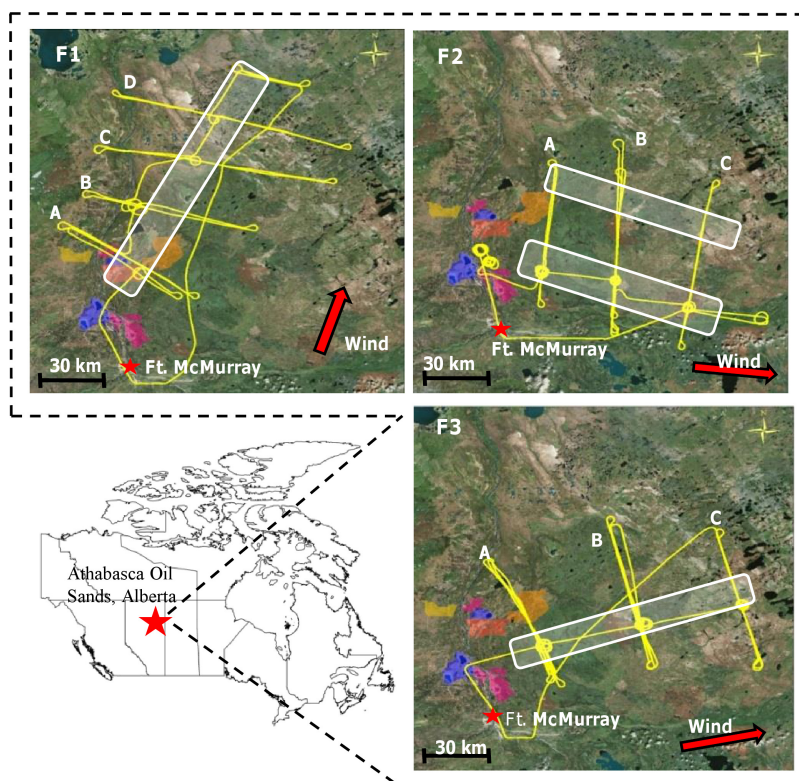
Additionally, the model incorporated SOA formation from all known SOA precursors²⁴ treating SOA formation in two separate volatility basis sets (VBSs) (see supplementary Methods). Following a previously described method²⁴, a four-bin VBS ($C^* = 1, 10, 100$ and $1,000 \mu g m^{-3}$) treated SOA formation from traditional volatile organic compounds (VOCs), while a second nine-bin VBS ($C^* = 10^{-2} - 10^6 \mu g m^{-3}$) treated SOA from SVOCs and IVOCs. The four-bin VBS was used for SOA from traditional VOCs including long-chain alkanes (ALK5 in SAPRC07), olefins (OLE1 and OLE2), aromatics (ARO1, ARO2, NAPTH and benzene), and biogenic compounds (ISOP, TERP and SESQ (isoprene, monoterpenes and sesquiterpenes))^{24,44}. The nine-bin VBS treated 'non-traditional' SOA formed from the oxidation of off-road diesel as well as bitumen vapours having a volatility distribution as shown in Fig. 4a at 20 °C. This volatility distribution was chosen to represent the emissions of these vapours at ambient temperature that would be expected for the first aircraft screen at ~600 m above ground, assuming that the open-pit mines are the largest contributor to emissions. A contribution by other processes at higher temperature is also possible. Total non-methane hydrocarbon (NMHC) mixing ratios in the plume were estimated based on the emission ratios of CO:NMHC from the heavy hauler diesel engines used in the Alberta OS facilities and the difference between CO in the plume and CO in the background (ΔCO). The emission ratios of SVOCs and IVOCs relative to total NMHC that were reported previously³⁹ for diesel engines were then applied to the total NMHC to give an estimate of the SVOCs and IVOCs in the plume. Pentadecane was used as a surrogate species for the SVOC and IVOC species from diesel emissions as suggested previously⁴⁴.

The model is configured in such a way that the initial reaction of a SOA precursor with OH (or O_3 in the case of ISOP, TERP, OLE1 and OLE2) leads to the formation of a number of less volatile gas-phase species. These less volatile gas-phase species are placed in volatility bins according to fitted chamber results⁴⁵. The species in each of the bins are then allowed to partition between the gas and particle phase in accordance with their temperature-dependent partitioning coefficients^{24,45}. To mimic aerosol ageing, the gas phase components in both the VOC SOA (V-SOA) and semi- and intermediate-volatility SOA (SI-SOA) VBS are aged as described previously²⁴. Specifically, traditional SOA in the V-SOA VBS is aged according to the Robinson *et al.* scheme⁴⁶, while SOA in the SI-SOA VBS is aged according to the more aggressive Grieshop scheme⁴⁷. The Robinson scheme used to age V-SOA adds 7.5% more mass to the SOA during oxidation

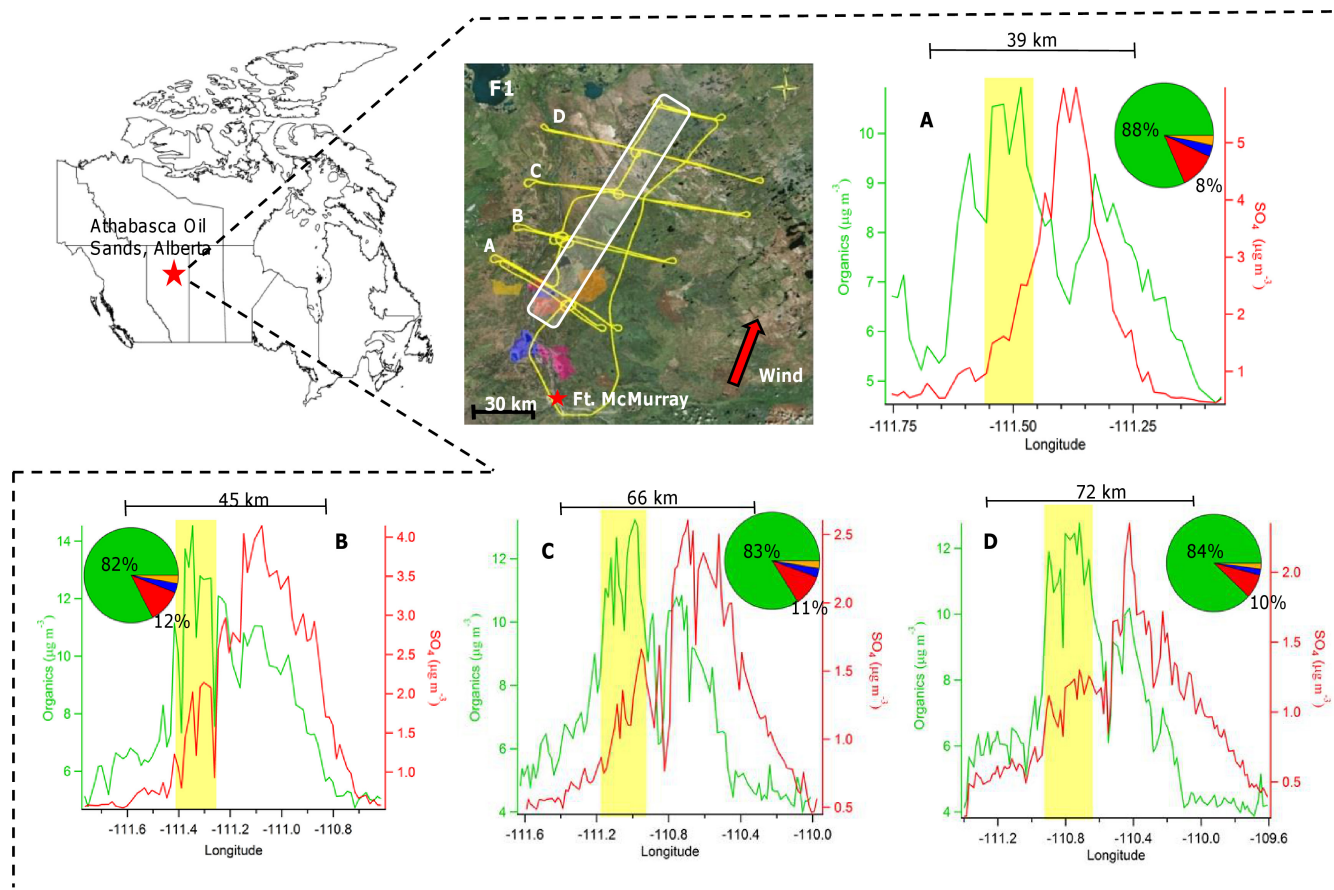
while moving the species to a volatility bin 10 times less volatile. The Grieshop scheme⁴⁷ that was used to age the SI-SOA adds 40% more mass per oxidation but shifts the species to a volatility bin 100 times less volatile. As the majority of the SOA formed in the V-SOA VBS is formed from anthropogenic precursors, V-SOA was aged at a rate of $1 \times 10^{-11} cm^3 molecule^{-1} s^{-1}$ (refs 48, 49). The SOA in the SI-SOA VBS was aged using a faster rate of $2 \times 10^{-11} cm^3 molecule^{-1} s^{-1}$ (ref. 24). The use of two separate ageing schemes for SOA formation is consistent with the expected differences between product distributions, molecular size and functional groups of different classes of precursor organic compounds. Such an approach has been used successfully on numerous occasions to match SOA observations (see Supplementary Methods). Further model runs were also performed to examine the sensitivity of the SOA formed from IVOCs to the oxidation scheme used (Extended Data Fig. 9 and Supplementary Methods). On the basis of these further model runs, the chosen base case conditions provide the best estimate of the SOA formation rate as it lies between the two upper and lower limits and is consistent with the scheme used in numerous regional air quality models that reasonably reproduce ambient forested and urban observations around the world.

The model output was compared with organic aerosol observations. While the HR-ToF-AMS effectively measures $PM_{1.0}$, the condensation of oxidized products will occur across the entire size distribution. Considerable coarse particle mass is observed during flight 1, probably originating from the large trucks during mining operations. Since the box-model output is a bulk SOA value (that is, size independent), the AMS-derived OA mass is further increased using the measured surface area ratio of $PM_{1.0}$ to $PM_{2.5}$, assuming that the condensation process is approximately proportional to surface area. This ratio, which ranged from ~1.3 to 1.1 from screen A to screen D, was multiplied by the AMS-measured OA, increasing the total OA by 10–30% for comparison to the model output.

30. DeCarlo, P. F. *et al.* Field-deployable, high-resolution, time-of-flight aerosol mass spectrometer. *Anal. Chem.* **78**, 8281–8289 (2006).
31. Moteki, N. & Kondo, Y. Dependence of laser-induced incandescence on physical properties of black carbon aerosols: measurements and theoretical interpretation. *Aerosol Sci. Technol.* **44**, 663–675 (2010).
32. Schwarz, J. P. *et al.* Single-particle measurements of midlatitude black carbon and light-scattering aerosols from the boundary layer to the lower stratosphere. *J. Geophys. Res.* **111**, D16207 (2006).
33. de Gouw, J. & Warneke, C. Measurements of volatile organic compounds in the earth's atmosphere using proton-transfer-reaction mass spectrometry. *Mass Spectrom. Rev.* **26**, 223–257 (2007).
34. Garratt, J. R. *The Atmospheric Boundary Layer* (Cambridge Univ. Press, 1994).
35. Vinuesa, J. F. & Galmarini, S. Turbulent dispersion of non-uniformly emitted passive tracers in the convective boundary layer. *Boundary-Layer Meteorol.* **133**, 1–16 (2009).
36. Carter, W. P. L. Development of a condensed SAPRC-07 chemical mechanism. *Atmos. Environ.* **44**, 5336–5345 (2010).
37. Chen, Y., Sexton, K. G., Jerry, R. E., Surratt, J. D. & Vizuete, W. Assessment of SAPRC07 with updated isoprene chemistry against outdoor chamber experiments. *Atmos. Environ.* **105**, 109–120 (2015).
38. Xie, Y. *et al.* Understanding the impact of recent advances in isoprene photooxidation on simulations of regional air quality. *Atmos. Chem. Phys.* **13**, 8439–8455 (2013).
39. de Gouw, J. *et al.* Sensitivity and specificity of atmospheric trace gas detection by proton-transfer-reaction mass spectrometry. *Int. J. Mass Spectrom.* **223–224**, 365–382 (2003).
40. Zhao, J. & Zhang, R. Proton transfer reaction rate constants between hydronium ion (H_3O^+) and volatile organic compounds. *Atmos. Environ.* **38**, 2177–2185 (2004).
41. Hakola, H. *et al.* Seasonal variation of mono- and sesquiterpene emission rates of Scots pine. *Biogeosciences* **3**, 93–101 (2006).
42. Helmig, D. *et al.* Sesquiterpene emissions from pine trees—identifications, emission rates and flux estimates for the contiguous United States. *Environ. Sci. Technol.* **41**, 1545–1553 (2007).
43. Ehn, M. *et al.* A large source of low-volatility secondary organic aerosol. *Nature* **506**, 476–479 (2014).
44. Jathar, S. H. *et al.* Unspecified organic emissions from combustion sources and their influence on the secondary organic aerosol budget in the United States. *Proc. Natl Acad. Sci. USA* **111**, 10473–10478 (2014).
45. Tsimpidi, A. P. *et al.* Evaluation of the volatility basis-set approach for the simulation of organic aerosol formation in the Mexico City metropolitan area. *Atmos. Chem. Phys.* **10**, 525–546 (2010).
46. Robinson, A. L. *et al.* Rethinking organic aerosols: semivolatile emissions and photochemical aging. *Science* **315**, 1259–1262 (2007).
47. Grieshop, A. P., Logue, J. M., Donahue, N. M. & Robinson, A. L. Laboratory investigation of photochemical oxidation of organic aerosol from wood fires 1: measurement and simulation of organic aerosol evolution. *Atmos. Chem. Phys.* **9**, 1263–1277 (2009).
48. Murphy, B. N., Donahue, N. M., Fountoukis, C. & Pandis, S. N. Simulating the oxygen content of ambient organic aerosol with the 2D volatility basis set. *Atmos. Chem. Phys.* **11**, 7859–7873 (2011).
49. Shrivastava, M. K., Lane, T. E., Donahue, N. M., Pandis, S. N. & Robinson, A. L. Effects of gas particle partitioning and aging of primary emissions on urban and regional organic aerosol concentrations. *J. Geophys. Res.* **113**, (2008).

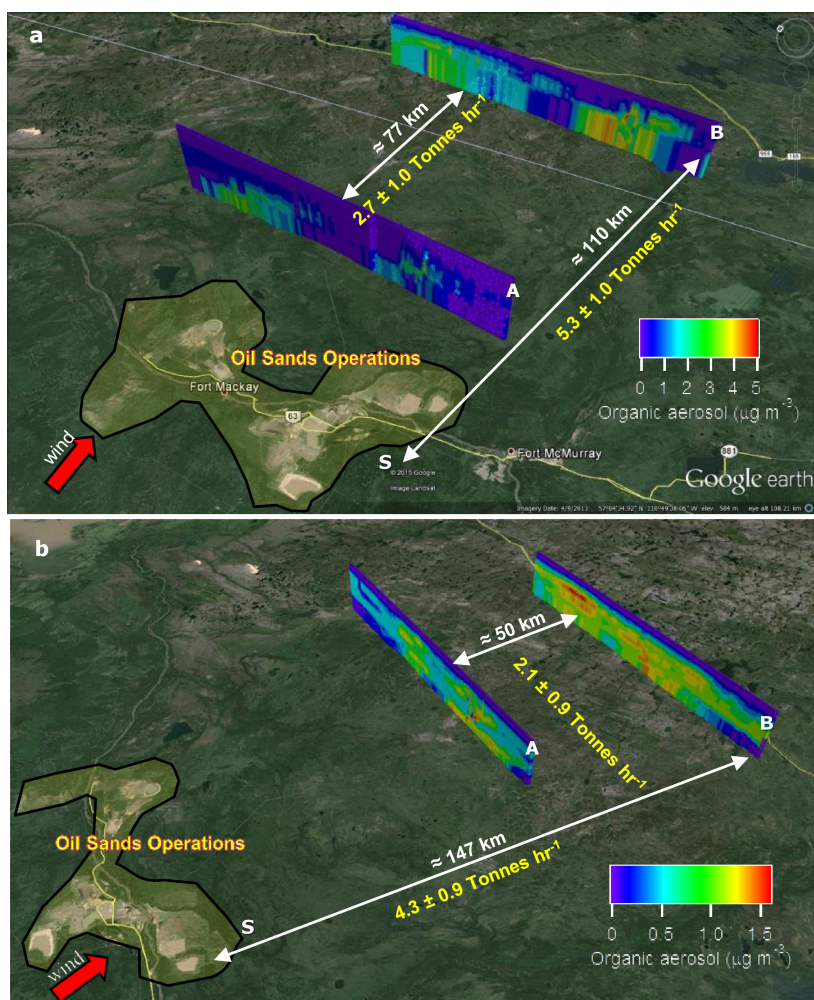


Extended Data Figure 1 | Flight tracks for the three transformation flights, F1, F2 and F3. The approximate locations of the major OS plumes studied in this work are shown as the white shaded boxes. Map data: Google, image Landsat, 2015.



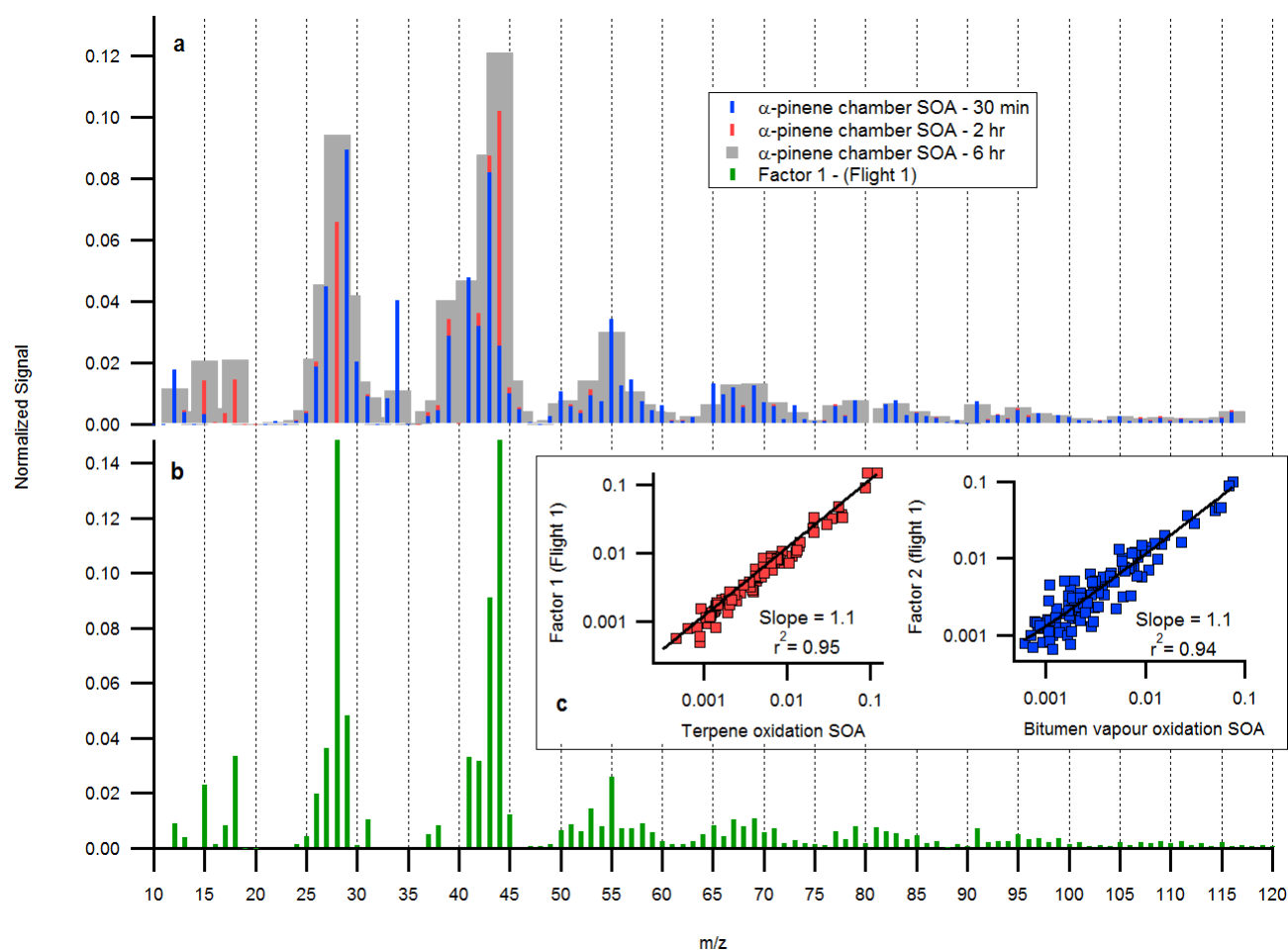
Extended Data Figure 2 | Measured organic and sulfate aerosol concentration during F1. Successive transects (labelled A, B, C and D) through the same major OS plumes at approximately 600 m altitude and 1 h apart in transit time. Inset pie plots show the mean relative mass fraction for organics (green), sulfate (red), nitrate (blue) and ammonium

(orange) during the yellow highlighted section. Organics dominate the aerosol mass throughout the flight; note the change in magnitude between the OA scale on the left and SO_4 scale on the right. Map data: Google, image Landsat, 2015.



Extended Data Figure 3 | OA mass screens used to estimate SOA production. a, b, OA mass screens for F2 (a) and F3 (b). The SOA production rate during these flights (~ 77 km and ~ 50 km between screens) is the sum of the differences in OA transfer rates between screens

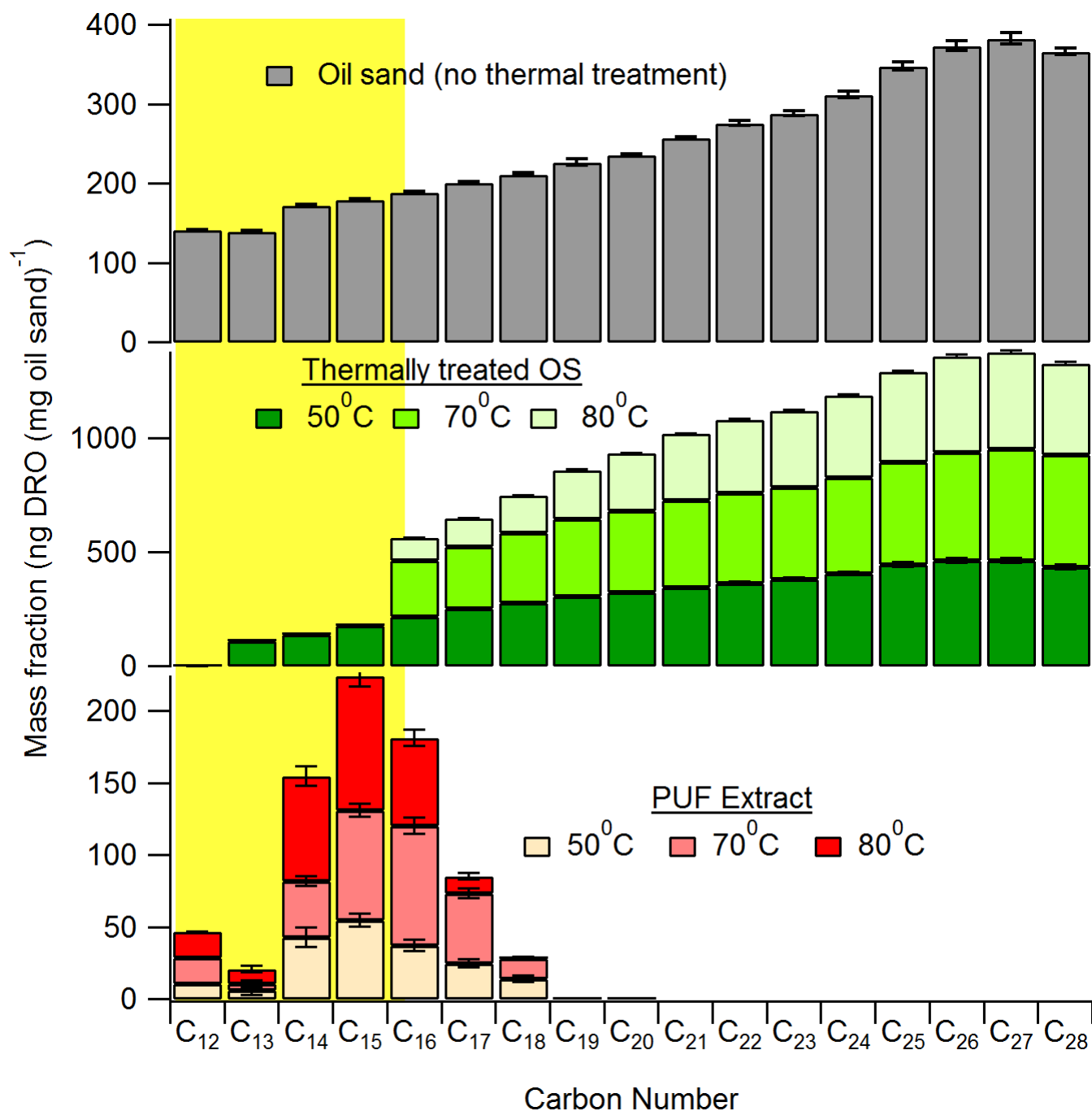
(that is, $2.7 \pm 1.0 \text{ t h}^{-1}$ and $2.1 \pm 0.9 \text{ t h}^{-1}$). The overall formation rate from the OS source region (S) is the integrated OA transfer rate through screen B ($5.3 \pm 1.0 \text{ t h}^{-1}$ and $4.3 \pm 0.9 \text{ t h}^{-1}$). Map data: Google, image Landsat, 2015.



Extended Data Figure 4 | PMF analysis results and comparisons.

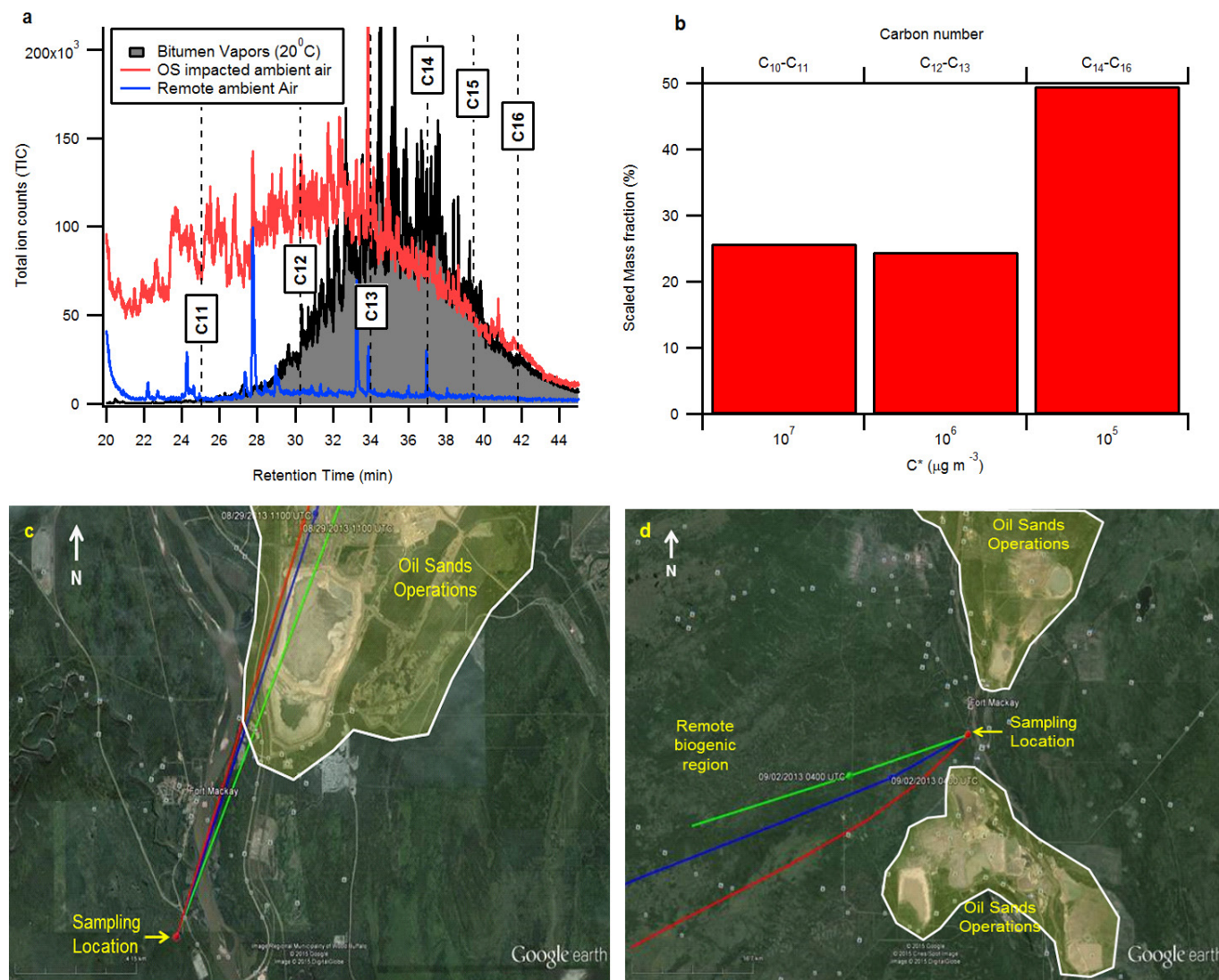
a, The OA AMS spectra from an α -pinene + OH radical smog chamber experiment as a function of photochemical ageing time in the chamber.
b, PMF factor 1 from F1. A high degree of similarity is observed between

these spectra after approximately 6 h of ageing in the chamber.
c, Correlations between PMF factors 1 and 2 and the corresponding smog chamber data (terpene oxidation and bitumen vapour oxidation SOA).



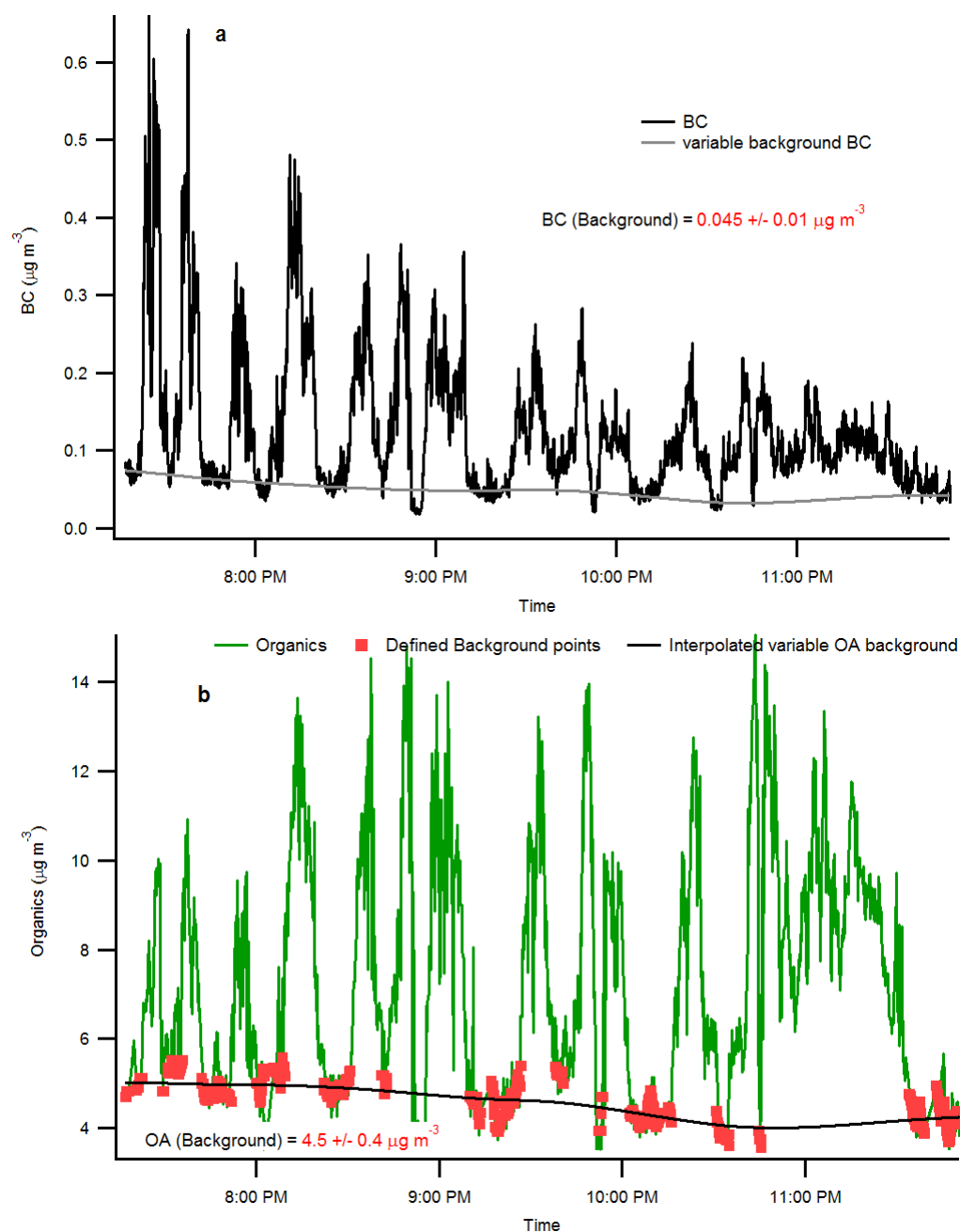
Extended Data Figure 5 | Bitumen volatility distributions. The volatility distribution (mass fraction) based on carbon number are for OS that was thermally treated. Volatile hydrocarbons are trapped on polyurethane foam (PUF) tubes at 50–80 °C (red). The volatility of the remaining bitumen material is shown in green (50–80 °C) and that of bitumen which

was solvent extracted from the sand without heating is shown in grey. Note the complete loss of hydrocarbons in the C₁₂–C₁₅ range upon heating (denoted in yellow). Data are stacked upon each other for clarity. Error bars represent the s.d. of $n = 3$ experiments.

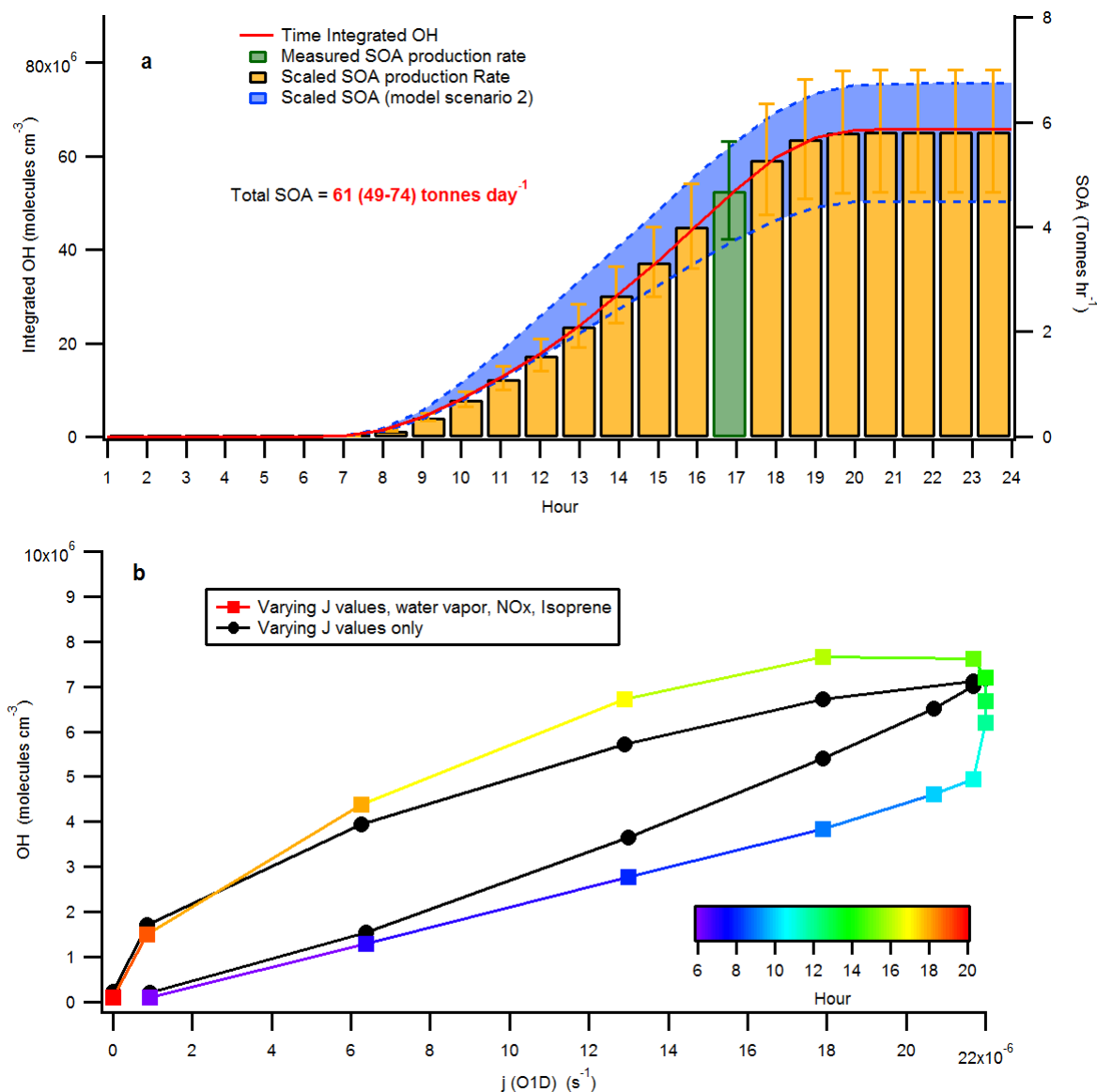


Extended Data Figure 6 | Bitumen-related IVOCs in ambient ground-based data. **a**, Total ion chromatogram from ambient sampling in the OS when impacted by forest-influenced air (blue) and OS-operations air (red). The bitumen vapour headspace chromatogram is also shown (black), demonstrating that a large fraction of the gaseous mass in OS-impacted air

has volatilities (C_{13} – C_{16} range) critical for SOA formation. **b**, Associated volatility distribution for OS-impacted air scaled by SOA yield¹¹. **c**, One-hour back trajectory for OS-impacted sample using the hybrid single particle Lagrangian integrated trajectory model (HYSPLIT). **d**, One-hour HYSPLIT back trajectory for forest-influenced sample.



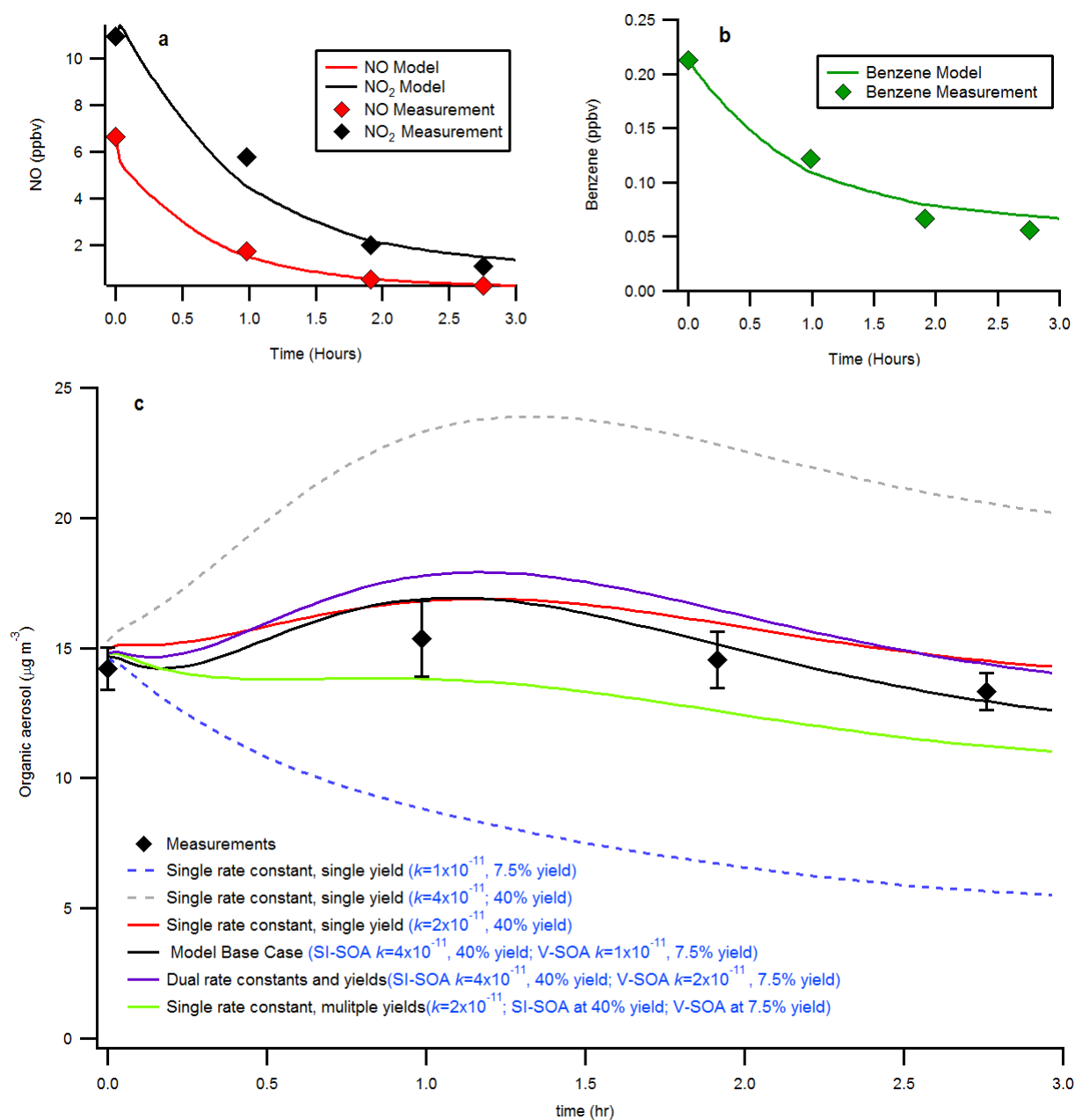
Extended Data Figure 7 | Background concentration time series. a, b, The BC (a) and OA (b) time series for F1 with associated interpolated backgrounds. The background variability contributed little uncertainty to the overall analysis of $\Delta\text{OA}/\Delta\text{BC}$ in Fig. 1.



Extended Data Figure 8 | SOA production rate extrapolation.

a, Measured SOA for F1 extrapolated to one photochemical day. Total SOA production is the sum of scaled hourly SOA production rates (orange; see Supplementary Methods). The blue region represents the same scaling performed where only photolysis rate constants are varied in the model.

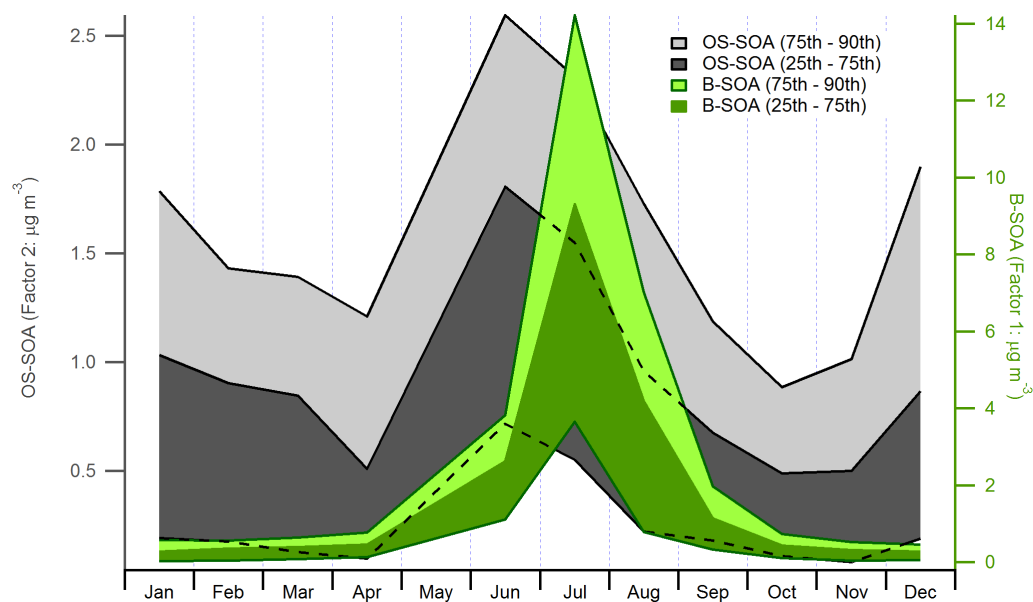
Error bars represent a range of SOA estimates assuming $\pm 20\%$ on the initial OA estimates via the TERRA algorithm. **b**, Modelled dependence of OH radical concentration on the ozone photolysis frequency ($JO1D$). Further varying initial conditions for NO_x, water vapour and isoprene in the model has a small effect on this relationship.



Extended Data Figure 9 | Box-model performance evaluation.

a, b, Measured and modelled gas-phase species during plume intercepts of F1, where only the initial conditions ($t = 0$) of the species are constrained by measurements. Good agreement between model and observation is achieved. **c,** Sensitivity of predicted SOA for F1 to changes in the oxidation

rate constant and yield (all other variables remain constant). Yield refers to the SOA mass yield during the oxidative ageing. Simulations using a single oxidative rate constant and yield represent upper and lower limits to SOA formation, while the base case simulation most closely resembles measurements. Error bars represent s.d. of the measured OA ($n = 7$).



Extended Data Figure 10 | PMF factors from ground-based data in the OS. PMF factors 1 (biogenic SOA (B-SOA)) and 2 (OS-SOA) from 1 year of ground-based data in the OS production region (monthly 25th to 90th percentiles shown, $n = 22,280$), indicating that factor 2 (using a collection

efficiency of 1) is derived from the oxidation of OS emissions all year long, while factor 1 is from oxidation of biogenic emissions (that is, summer peak only).

Experimental determination of the electrical resistivity of iron at Earth's core conditions

Kenji Ohta¹, Yasuhiro Kuwayama², Kei Hirose^{3,4}, Katsuya Shimizu⁵ & Yasuo Ohishi⁶

Earth continuously generates a dipole magnetic field in its convecting liquid outer core by a self-sustained dynamo action. Metallic iron is a dominant component of the outer core, so its electrical and thermal conductivity controls the dynamics and thermal evolution of Earth's core¹. However, in spite of extensive research, the transport properties of iron under core conditions are still controversial^{2–9}. Since free electrons are a primary carrier of both electric current and heat, the electron scattering mechanism in iron under high pressure and temperature holds the key to understanding the transport properties of planetary cores. Here we measure the electrical resistivity (the reciprocal of electrical conductivity) of iron at the high temperatures (up to 4,500 kelvin) and pressures (megabars) of Earth's core in a laser-heated diamond-anvil cell. The value measured for the resistivity of iron is even lower than the value extrapolated from high-pressure, low-temperature data using the Bloch–Grüneisen law, which considers only the electron–phonon scattering. This shows that the iron resistivity is strongly suppressed by the resistivity saturation effect at high temperatures. The low electrical resistivity of iron indicates the high thermal conductivity of Earth's core, suggesting rapid core cooling and a young inner core less than 0.7 billion years old¹⁰. Therefore, an abrupt increase in palaeomagnetic field intensity around 1.3 billion years ago¹¹ may not be related to the birth of the inner core.

Extensive efforts have been made to measure the electrical resistivity of iron at high pressure since the earliest high-pressure mineral physics experiments¹², but its direct measurement under the conditions of Earth's core is still challenging. Traditionally, the core resistivity was estimated to be 200–500 $\mu\Omega$ cm by a combination of static low-pressure, low-temperature measurements and shock compression data^{2,3}. However, measurements under shock-wave compression potentially overestimate the resistivity owing to defect production during shock impact¹³. Recent density functional theory calculations^{5,6,14,15} predicted the core resistivity to be 20%–50% of these conventional estimates. The static high-pressure, low-temperature experiments^{7,9} also suggested a low core resistivity because the resistivity saturates at high temperature. These values suggest that Earth's core has been cooling rapidly, which implies a young inner core and high initial core and mantle temperatures^{10,16}. However, the resistivity saturation and the resulting low electrical resistivity of Earth's core have not been verified by experiments at the relevant high-pressure, high-temperature conditions.

We made use of advanced experimental techniques, including the shaping of the sample and electrode composites, to measure the electrical resistivity ρ of iron at ultrahigh pressure and temperature in a laser-heated diamond-anvil cell (DAC) (Extended Data Fig. 1). We first examined the temperature response of the resistivity at 26 GPa up to 2,610 K (Fig. 1a). Kinks in the temperature–resistivity curve are associated with phase changes from ϵ to γ and from γ to liquid; these phase changes were confirmed by concurrent synchrotron X-ray

diffraction measurements (Extended Data Fig. 2). Our data at 26 GPa is in broad agreement with an earlier report at 15 GPa in a multi-anvil apparatus¹⁷. We observed similar behaviour at 51 GPa up to 2,880 K in a separate run (Fig. 1b). A resistivity jump of about +20% upon melting was observed in these two runs, although it may include some

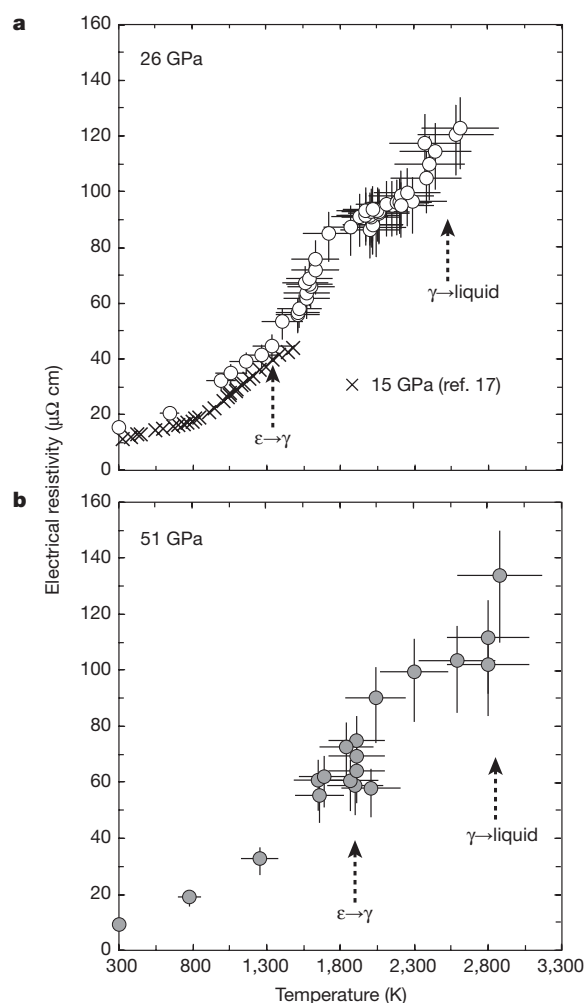


Figure 1 | Change in electrical resistivity of iron upon phase transitions. **a**, Measurements from this study taken at 26 GPa up to 2,610 K (open circles) are compared with previous measurements in a multi-anvil press at 15 GPa (ref. 17) (crosses). **b**, Measurements from this study taken at 51 GPa up to 2,880 K (grey circles). Error bars reflect the uncertainty in the high-pressure, high-temperature resistivity obtained (see Methods) and 1σ of the measured temperature variations.

¹Department of Earth and Planetary Sciences, Tokyo Institute of Technology, 2-12-1 Ookayama, Meguro, Tokyo 152-8551, Japan. ²Geodynamics Research Center, Ehime University, 2-5 Bunkyo-cho, Matsuyama, Ehime 790-8577, Japan. ³Earth-Life Science Institute, Tokyo Institute of Technology, 2-12-1 Ookayama, Meguro, Tokyo 152-8550, Japan. ⁴Laboratory of Ocean-Earth Life Evolution Research, Japan Agency for Marine-Earth Science and Technology, 2-15 Natsushima-cho, Yokosuka, Kanagawa 237-0061, Japan. ⁵Center for Science and Technology under Extreme Conditions, Osaka University, 1-3 Machikaneyama, Toyonaka, Osaka 560-8531, Japan. ⁶Japan Synchrotron Radiation Research Institute, 1-1-1 Koto, Sayo, Hyogo 679-5198, Japan.

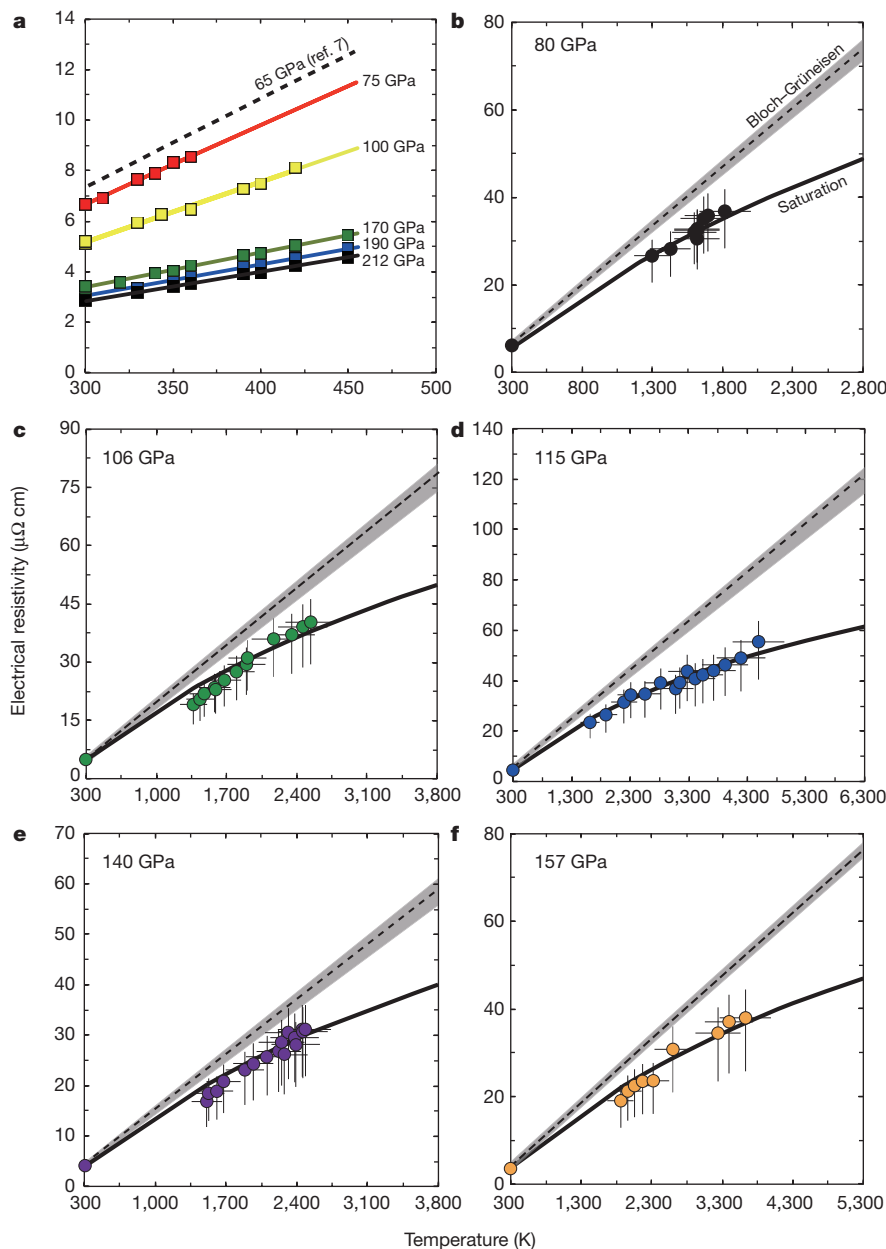


Figure 2 | Change in the resistivity of ϵ iron with increasing temperature.

a, Temperature dependence of the resistivity up to 450 K at 75–212 GPa fitted by the Bloch–Grüneisen formula (coloured lines). Similar results at 65 GPa from ref. 7 are also shown. The resistivity value includes the uncertainty (not shown) derived from that in the reference high-pressure, room-temperature resistivity data⁷ (Extended Data Fig. 4), but the slope is obtained from the change in resistance, which was measured with very small errors (see Methods). **b–f**, Electrical resistivity measured at 80 GPa

up to 1,820 K (**b**), at 106 GPa up to 2,540 K (**c**), at 115 GPa up to 4,490 K (**d**), at 140 GPa up to 2,490 K (**e**), and at 157 GPa up to 3,630 K (**f**). The resistivity measured at high pressure and high temperature in this study is lower than the prediction by the Bloch–Grüneisen formula, ρ_{BG} (dashed lines with grey uncertainty band) with parameters obtained from the high-pressure, low-temperature measurements in **a**. Such a low resistivity can be accounted for by the effect of resistivity saturation at high temperature (solid curves). Error bars are as in Fig. 1.

uncertainty related to the measurement of liquid iron, which is difficult owing to a change in sample geometry at melting. Previous experiments performed below 7 GPa reported that the resistivity of iron increased by 5%–9% upon melting^{18–20}, while a recent theoretical study showed a 13%–20% resistivity difference between solid iron (the ϵ phase) and liquid iron at 330 GPa (ref. 15).

We carried out more experiments both in a muffle furnace up to 450 K (Fig. 2a) and in a laser-heated DAC up to 4,500 K (Fig. 2b–f) at a higher pressure range in which only the ϵ phase was found (Extended Data Fig. 3). The former measurements indicated that the resistivity increased linearly with increasing temperature. Such a temperature–resistivity slope below 450 K is expressed by the Bloch–Grüneisen law:

$$\rho_{BG}(V, T) = D(V) \left(\frac{T}{\Theta_D(V)} \right)^n \int_0^{\frac{\Theta_D(V)}{T}} dz \frac{z^n}{(\exp(z) - 1)(1 - \exp(-z))} \quad (1)$$

in which both the Debye temperature Θ_D and the volume V are available from the literature²¹. n is an integer that depends upon the interaction of free electrons. The expression of the present data using equation (1) yields a material constant $D(V)$ and n value at each pressure (Table 1). We found that the temperature T dependence became weaker with each pressure increment, corresponding to a reduction in n from 5.5 to 1.5, which is reasonably consistent with previous results^{2,3,8}. This suggests

Table 1 | Parameters for the Bloch–Grüneisen formula

P (GPa)	Θ_D (K)	D (V)	n
65*	610	92.6(1)	5.9
75	626	79.6(5)	5.5
100	677	33.3(4)	3.3
170	783	8.1(1)	1.8
190	809	5.6(1)	1.6
212	836	4.5(0)	1.5

Θ_D was calculated from the equation of state of ε iron²¹. Fitting errors in n are smaller than the first decimal place.

*From ref. 7.

that increasing pressure diminishes impedance against free electron migration in iron.

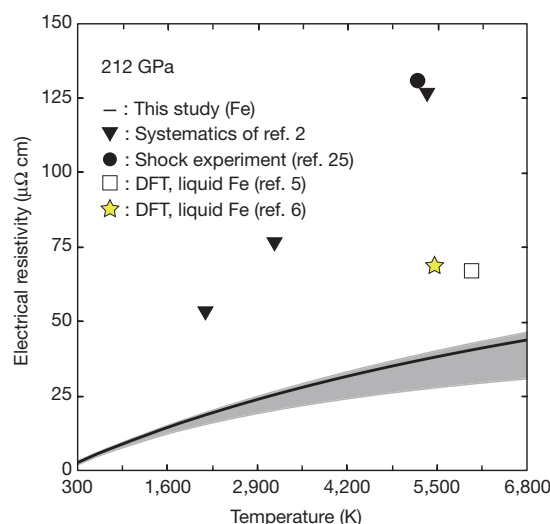
Using the $D(V)$ and n values obtained from the high-pressure, low-temperature measurements, the Bloch–Grüneisen formula predicts the resistivity of iron at high-pressure, high-temperature conditions. However, the present experiments, performed up to 4,500 K in the range 80–157 GPa, demonstrate that the measured resistivity is certainly lower than the value predicted by the Bloch–Grüneisen law (ρ_{BG}) even when we consider all possible error sources (see Methods). In principle, electron–electron scattering may be important at very high temperatures. However, we observe no sign of resistivity enhancement due to the electron–electron scattering with increasing temperature, at least up to 4,500 K (Fig. 2b–f). The low resistivity of ε iron observed in this study may be attributed to the well known²² effect of resistivity saturation at high temperature⁷ (in which the resistivity increase is suppressed at high temperature). The electrical resistivity of metal asymptotically approaches the Ioffe–Regel value (that is, saturation resistivity, ρ_{sat}) when the mean-free path of free electrons becomes comparable to the interatomic distance²². Wiesmann *et al.*²³ proposed an empirical description of the resistivity saturation in a simple shunt resistor model that can be applied to a variety of metals:

$$\frac{1}{\rho_{BG+sat}} = \frac{1}{\rho_{BG}} + \frac{1}{\rho_{sat}} \quad (2)$$

The present temperature–resistivity data at each pressure is well explained by the shunt resistor model of equation (2) with a reasonable value of ρ_{sat} (Fig. 2b–f). The saturation resistivity ρ_{sat} should decrease under compression, because it depends on the interatomic spacing. Indeed, the fitting results show that ρ_{sat} diminishes from 142 $\mu\Omega$ cm at 80 GPa to 122 $\mu\Omega$ cm at 157 GPa, although the uncertainties are somewhat large. These values are in good accordance with the value measured at 1 bar (168 $\mu\Omega$ cm) (ref. 24) (Extended Data Table 1). Thus, the present results up to 4,500 K demonstrate the effect of resistivity saturation, which defines the upper bound for the resistivity and leads to the low electrical resistivity (high thermal conductivity) of Earth's core.

The resistivity (ρ_{BG+sat}) of ε iron at 212 GPa is shown as a function of temperature and compared to previous estimates^{2,5,6,25} in Fig. 3. The temperature–resistivity curve is calculated using the shunt resistor model (equation (2)), with $D(V) = 4.5 \mu\Omega$ cm, $n = 1.5$, and $\rho_{sat} = 116_{-60}^{+18} \mu\Omega$ cm, which were measured at 212 GPa in this study (Table 1) or estimated by linear extrapolation of the ρ_{sat} versus $(V/V_0)^{1/3}$ relation (Extended Data Table 1). Our estimate gives the lowest value because of the effect of resistivity saturation, which was not considered in earlier models of core conductivity^{2–3}. The values of resistivity of liquid iron predicted by de Koker *et al.*⁵ and Pozzo *et al.*⁶ agree well with the present value for solid iron when we consider a $\sim 20\%$ increase upon melting (Fig. 3).

The present data demonstrate the electrical resistivity of iron to be $40.4_{-9.7}^{+6.5} \mu\Omega$ cm at 140 GPa and 3,750 K (Fig. 2e), close to the core–mantle boundary conditions. It corresponds to the electronic thermal conductivity $\kappa_{el} = 226_{-31}^{+71} \text{ W m}^{-1} \text{ K}^{-1}$ when the Wiedemann–Franz relation ($\kappa_{el} = L_0 T / \rho$, for the ideal Lorenz number $L_0 = 2.44 \times 10^{-8} \text{ W } \Omega \text{ K}^{-2}$) is

**Figure 3 | Iron resistivity at 212 GPa and high temperatures.**

Comparison of the present results of iron resistivity at 212 GPa (black solid curve with grey uncertainty band) with previous modelling² (triangles), shock compression study²⁵ (circle), density functional theory calculations (square⁵, star⁶).

applied. The Lorenz number for iron at core–mantle boundary conditions recently computed by theoretical studies^{5,14} shows less than $+3\%/-6\%$ difference from the ideal L_0 value.

Since Earth's core contains some nickel and light elements in addition to iron, we consider the effect of such impurity elements. The impurity resistivity in iron has been measured for silicon^{7,8} and nickel⁹ at high-pressure, room-temperature (300 K) conditions (see Methods). From Matthiessen's rule, the resistivity of solid $\text{Fe}_{77.5}\text{Ni}_{10}\text{Si}_{22.5}$, a possible outer core composition inferred from its density²⁶, is calculated to be $86.9_{-21.6}^{+15.4} \mu\Omega$ cm at 140 GPa and 3,750 K, considering the saturation effect. When the 20% resistivity increase upon melting is taken into account, we obtain $104_{-26}^{+18} \mu\Omega$ cm and thus a thermal conductivity of $88_{-13}^{+29} \text{ W m}^{-1} \text{ K}^{-1}$ for liquid $\text{Fe}_{67.5}\text{Ni}_{10}\text{Si}_{22.5}$.

Recent modelling of core thermal evolution¹⁰ using the value for thermal conductivity of solid $\text{Fe}_{77.5}\text{Si}_{22.5}$ obtained by Gomi *et al.*⁷ demonstrates that Earth's core has been cooling quickly and that the inner core is less than 0.7 billion years old. Our estimate of core thermal conductivity at the core–mantle boundary coincides with the $90 \text{ W m}^{-1} \text{ K}^{-1}$ reported by Gomi *et al.*⁷, although they assumed a higher saturation resistivity and did not consider the effect of melting. As a consequence, the present study supports a young age for the inner core.

Biggin *et al.*¹¹ found that the intensity and variability of the geomagnetic field were enhanced around 1.3 billion years ago and attributed this enhancement to the beginning of nucleation of the solid inner core. Such a remarkable change in palaeomagnetic field, however, may have been caused by another effect such as a change in spatial variation in the core–mantle boundary heat flux²⁷. In addition, Earth's magnetic field has been present over a long geological time (since 4.2 billion years ago²⁸) and it has been assumed that the geomagnetic field was induced by thermal convection in the core before the onset of inner-core crystallization. The high thermal conductivity obtained in this study, however, suggests that this is not the case because otherwise the core must have been unrealistically hot in the early history of Earth¹⁰. Alternatively, core convection in the absence of an inner core might have been driven or assisted by libration, precession and tides²⁹. The precipitation of magnesium-bearing minerals from the core could also have provided an additional energy source to promote core convection³⁰.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 1 May 2015; accepted 17 March 2016.

1. Hirose, K., Labrosse, S. & Hernlund, J. Composition and state of the core. *Annu. Rev. Earth Planet. Sci.* **41**, 657–691 (2013).
2. Stacey, F. & Anderson, O. Electrical and thermal conductivities of Fe–Ni–Si alloy under core conditions. *Phys. Earth Planet. Inter.* **124**, 153–162 (2001).
3. Stacey, F. & Loper, D. A revised estimate of the conductivity of iron alloy at high pressure and implications for the core energy balance. *Phys. Earth Planet. Inter.* **161**, 13–18 (2007).
4. Sha, X. & Cohen, R. E. First-principles studies of electrical resistivity of iron under pressure. *J. Phys. Condens. Matter* **23**, 075401 (2011).
5. de Koker, N., Steinle-Neumann, G. & Vlcek, V. Electrical resistivity and thermal conductivity of liquid Fe alloys at high P and T, and heat flux in Earth's core. *Proc. Natl Acad. Sci. USA* **109**, 4070–4073 (2012).
6. Pozzo, M., Davies, C., Gubbins, D. & Alfè, D. Thermal and electrical conductivity of iron at Earth's core conditions. *Nature* **485**, 355–358 (2012).
7. Gomi, H. *et al.* The high conductivity of iron and thermal evolution of the Earth's core. *Phys. Earth Planet. Inter.* **224**, 88–103 (2013).
8. Seagle, C., Cottrell, E., Fei, Y., Hummer, D. & Prakapenka, V. Electrical and thermal transport properties of iron and iron–silicon alloy at high pressure. *Geophys. Res. Lett.* **40**, 5377–5381 (2013).
9. Gomi, H. & Hirose, K. Electrical resistivity and thermal conductivity of hcp Fe–Ni alloys under high pressure: implications for thermal convection in the Earth's core. *Phys. Earth Planet. Inter.* **247**, 2–10 (2015).
10. Labrosse, S. Thermal evolution of the core with a high thermal conductivity. *Phys. Earth Planet. Inter.* **247**, 36–55 (2015).
11. Biggin, A. *et al.* Palaeomagnetic field intensity variations suggest Mesoproterozoic inner-core nucleation. *Nature* **526**, 245–248 (2015).
12. Bridgman, P. W. The resistance of 72 elements, alloys and compounds to 100,000 kg/cm². *Proc. Am. Acad. Arts Sci.* **81**, 165–251 (1952).
13. Murr, L., Inal, O. & Morales, A. Vacancies and vacancy clusters in shock-loaded molybdenum: direct observations by transmission electron and field-ion microscopy. *Appl. Phys. Lett.* **28**, 432–434 (1976).
14. Pozzo, M., Davies, C., Gubbins, D. & Alfè, D. Transport properties for liquid silicon–oxygen–iron mixtures at Earth's core conditions. *Phys. Rev. B* **87**, 014110 (2013).
15. Pozzo, M., Davies, C., Gubbins, D. & Alfè, D. Thermal and electrical conductivity of solid iron and iron–silicon mixtures at Earth's core conditions. *Earth Planet. Sci. Lett.* **393**, 159–164 (2014).
16. Davies, C., Pozzo, M., Gubbins, D. & Alfè, D. Constraints from material properties on the dynamics and evolution of Earth's core. *Nature Geosci.* **8**, 678–685 (2015).
17. Deng, L., Seagle, C., Fei, Y. & Shahar, A. High pressure and temperature electrical resistivity of iron and implications for planetary cores. *Geophys. Res. Lett.* **40**, 33–37 (2013).
18. Powell, R. W. The electrical resistivity of liquid iron. *Phil. Mag.* **44**, 772–775 (1953).
19. Van Zylteld, J. B. Electrical resistivities of liquid transition metals. *J. Phys. Colloq.* **41** (C8), 503–506 (1980).
20. Secco, R. & Schloessin, H. The electrical resistivity of solid and liquid Fe at pressures up to 7 GPa. *J. Geophys. Res.* **94**, 5887–5894 (1989).
21. Dewaele, A. *et al.* Quasihydrostatic equation of state of iron above 2 Mbar. *Phys. Rev. Lett.* **97**, 215504 (2006).
22. Gunnarsson, O., Calandra, M. & Han, J. E. Saturation of electrical resistivity. *Rev. Mod. Phys.* **75**, 1085–1099 (2003).
23. Wiesmann, H. *et al.* Simple model for characterizing the electrical resistivity in A-15 superconductors. *Phys. Rev. Lett.* **38**, 782–785 (1977).
24. Bohnenkamp, U., Sandström, R. & Grimvall, G. Electrical resistivity of steels and face-centered-cubic iron. *J. Appl. Phys.* **92**, 4402–4407 (2002).
25. Bi, Y., Tan, H. & Jing, F. Electrical conductivity of iron under shock compression up to 200 GPa. *J. Phys. Condens. Matter* **14**, 10849–10854 (2002).
26. Sata, N. *et al.* Compression of FeSi, Fe₃C, Fe_{0.95}O, and FeS under the core pressures and implication for light element in the Earth's core. *J. Geophys. Res.* **115**, B09204 (2010).
27. Olson, P., Deguen, R., Hinnov, L. A. & Zhong, S. Controls on geomagnetic reversals and core evolution by mantle convection in the Phanerozoic. *Phys. Earth Planet. Sci.* **214**, 87–103 (2013).
28. Tarduno, J. A., Cottrell, R. D., Davis, W. J., Nimmo, F. & Bono, R. K. A Hadean to Paleoproterozoic geodynamo recorded by single zircon crystals. *Science* **349**, 521–524 (2015).
29. Le Bars, M., Cébron, D. & Le Gal, P. Flows driven by libration, precession, and tides. *Annu. Rev. Fluid Mech.* **47**, 163–193 (2015).
30. O'Rourke, J. G. & Stevenson, D. J. Powering Earth's dynamo with magnesium precipitation from the core. *Nature* **529**, 387–389 (2016).

Acknowledgements We thank H. Gomi for discussions and assistance with experiments. H. Ichikawa helped with the temperature distribution calculations. High-pressure experiments were conducted at BL10XU, SPring-8 (proposal numbers 2012B1131, 2012B1212, 2013B0080, 2014A0080, and 2014B0080).

Author Contributions K.O. designed the project. Y.K., K.H., K.S. and Y.O. supported the experiments. The manuscript was written by K.O. and K.H. and reviewed by all authors.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to K.O. (k-ohta@geo.titech.ac.jp).

METHODS

High-pressure, high-temperature resistance measurements. We used symmetric-type DACs with 40- μm , 90- μm , 120- μm , 150- μm , and 300- μm culet diamond anvils to generate high pressures. The sample was iron foil (99.99% purity), the same as that used in Gomi *et al.*⁷. The iron foil was shaped into a single member with four probes by using a focused ion beam apparatus (JEOL JIB-4000 and FEI Versa 3D) (Extended Data Fig. 1a). This shaping technique enables us to prepare samples with uniform geometry corresponding to each anvil culet size. The pressure medium was fine-grained SiO_2 glass, KCl or Al_2O_3 , which also acted as a thermal insulator during laser heating. They were loaded into a sample chamber at the centre of an insulated gasket consisting of rhenium and cubic boron nitride + epoxy powder. Four electrical leads made of platinum were connected to each iron lead outside the sample chamber (Extended Data Fig. 1b). Electrical resistance of an iron sample was measured using the four-terminal method to eliminate the large errors associated with lead resistance, using a Multimeter (Keithley 2000) or a SourceMeter (Keithley 2450) under a constant direct current of 100 mA.

Heating experiments were conducted in a muffle furnace up to 450 K (Fig. 2a). Sample temperature was monitored by a thermocouple. We obtained pressure at room temperature based on the Raman spectrum of a diamond anvil³¹. We also measured the temperature response of the electrical resistance of iron at high pressure in a laser-heated DAC (Figs 1 and 2b–f). The sample was heated with a couple of 100-W single-mode Yb fibre lasers using a double-side heating system at BL10XU, SPring-8. The laser-heated spot was 40 μm across, which was larger than the distance between two potential leads (Extended Data Fig. 1c). Temperature was obtained by a spectroradiometric method³². The variations in temperature within an area of resistance measurement were less than $\pm 10\%$. Since temperature heterogeneity in a heated area of the iron sample generates thermoelectric power, we measured the voltages of iron samples with passing direct current from both current directions (I_+ to I_- and I_- to I_+) and averaged these two voltage values to eliminate the effect of thermoelectric power (Extended Data Fig. 1b, c).

Concurrently with all high-pressure, high-temperature resistance measurements, we performed synchrotron X-ray diffraction (XRD) measurements at BL10XU, SPring-8 (Extended Data Fig. 2). Pressure was calculated from the unit-cell volumes of ϵ and γ iron and their pressure–volume–temperature equations of state^{21,33} (Extended Data Fig. 3). The pressures given in Figs 1 and 2 are those at 300 K. They increased slightly with increasing temperature, and the pressure increase itself diminished the resistivity, but it was corrected to examine the effect of temperature on iron resistivity at a given pressure (see below).

Estimation of resistivity. The electrical resistivity of iron at high pressure and high temperature was calculated from the ratio of resistance measured at high pressure and high temperature to that measured at high pressure and room temperature multiplied by the high-pressure, room-temperature resistivity of ϵ iron previously obtained by Gomi *et al.*⁷. Extended Data Fig. 4 compiles previously reported high-pressure, room-temperature resistivity values for ϵ iron^{4,7,8,34}. Some difference is found at relatively low pressures, but all of the experimental and theoretical determinations are consistent with each other above ~ 80 GPa. The present XRD analysis showed that sample pressure increased on laser heating in a DAC. The resistivity measured at high pressure and high temperature was corrected for the effect of such a pressure increase. To examine the temperature response of iron resistivity at constant pressure, we first estimated the relative reduction in iron resistivity with a certain pressure increment based on high-pressure, room-temperature resistivity data⁷ and then corrected the measured high-pressure, high-temperature resistivity value with that rate of reduction.

From the high-pressure, low-temperature (up to 450 K) data, we obtained the slope of the temperature–resistivity relation in a wide pressure range (65–212 GPa) and converted them into $D(V)$ and n parameters in the Bloch–Grüneisen formula, which can be compared with previous estimates^{2,7,8}. We assumed $D(V)$ and n to be temperature-independent. The saturation resistivity was calculated up to 157 GPa on the basis of the measured high-pressure, high-temperature resistivity data and the calculated Bloch–Grüneisen values, taking all possible error sources into account (Extended Data Table 1). Using these Bloch–Grüneisen values and the saturation resistivity at each pressure, we calculated the resistivities of ϵ iron up to $> 6,000$ K at 212 GPa (Fig. 3) and of $\text{Fe}_{67.5}\text{Ni}_{10}\text{Si}_{22.5}$ to 3,750 K at 140 GPa on the basis of a shunt resistor model (see equation (2)).

The uncertainty in the present high-pressure, high-temperature electrical resistivity measurement is derived from the following uncertainties: (1) in the measured resistance (only 0.012% error when using Keithley 2000 and 2450); (2) caused by volume expansion upon heating; and (3) in the reference high-pressure, room-temperature resistivity of ϵ iron determined by Gomi *et al.*⁷. We confirmed in each experiment that the geometry of the iron sample did not change before and after laser heating under optical microscope (Extended Data Fig. 1b). Indeed, iron resistance at 300 K before heating was identical to that after heating. XRD patterns collected before and after the high-pressure, high-temperature resistivity

measurements indicate that no chemical reaction occurred during laser heating (Extended Data Fig. 2b). The volume expansion observed by XRD data taken at high pressure and high temperature caused an underestimation of resistivity by $< 0.3\%$. The error in the reference high-pressure, room-temperature data is the main source of uncertainty; the resistivity of iron is $3.9^{+0.6}_{-0.9} \mu\Omega \text{ cm}$ at 140 GPa and 300 K (Extended Data Fig. 4). The error bars for the present high-pressure, high-temperature resistivity data in Figs 1 and 2 include all of these uncertainties.

Temperature distribution within a laser-heated sample. We simulated a temperature distribution within a laser-heated sample under high-pressure, high-temperature conditions (Extended Data Fig. 5). A steady-state heat conduction equation was employed:

$$\nabla(\kappa(T)\nabla T(x, y, z)) + A(x, y, z) = 0 \quad (3)$$

where $\kappa(T)$ is thermal conductivity and $A(x, y, z)$ is the energy flux from a laser, considering the heat balance at the sample surface.

Here we calculated the temperature distribution in an iron sample sandwiched by thermal insulation layers of Al_2O_3 at 115 GPa and 4,500 K, the conditions being the same as that for resistance measurement (Fig. 2d). The width, length (separation between two potential leads; V_+ and V_-), and thickness of the iron sample were 10 μm , 2 μm and 1 μm , respectively (the thickness was measured with a microprobe after decompression to 1 bar) (Extended Data Fig. 1b). Other dimensions were: thickness of Al_2O_3 layers, 2 μm ; diameter of a sample chamber, 40 μm ; culet and height of diamond anvils, 120 μm and 2 mm, respectively; laser beam size, 30 μm at full width of half maximum. We used the value for the temperature-dependent thermal conductivity of iron at 115 GPa given in Fig. 2d. The low thermal conductivity of iron, one-third of the present value and of a similar estimate by ref. 2, was also considered. The conductivities of the surrounding materials that we used were: single-crystal diamond anvil, $500 \text{ W m}^{-1} \text{ K}^{-1}$ (ref. 35); Al_2O_3 thermal insulator, $2.5 \text{ W m}^{-1} \text{ K}^{-1}$ (ref. 36) (for the very fine powdered Al_2O_3 used in the present experiments, the thermal resistance at the grain boundary is very large, making the bulk thermal conductivity much lower than that of a single crystal). Since there is no report for the thermal conductivity of the cubic boron nitride + epoxy mixture, we assumed it to be similar to the value of $60 \text{ W m}^{-1} \text{ K}^{-1}$ for polycrystalline hexagonal boron nitride at 1 bar (ref. 37). Here we did not consider pressure and temperature effects on the thermal conductivity of these surrounding materials. However, the thermal conductivity of these insulators increases with increasing pressure, while it decreases with increasing temperature, and these effects are cancelled out at high-pressure, high-temperature conditions^{38,39}.

With 4,500 K as a peak temperature at 115 GPa, our simulations show that the maximum temperature difference in the iron sample within an area for resistance measurement is about 200 K (Extended Data Fig. 5), smaller than the uncertainty in temperature determination. Even when we use low thermal conductivity for iron, the temperature difference is around 300 K. Such a temperature difference of 200–300 K changes the resistivity of iron by only $\sim 3\%$ at 115 GPa at around 4,000 K. These heat conduction calculations indicate that the observed low iron resistivity was not derived from a strong temperature gradient within a sample but is the intrinsic nature of the electrical property of iron.

Impurity effect on electrical resistivity of iron. Earth's core consists not only of iron but also of nickel and light element(s) such as silicon, oxygen, sulphur, carbon and hydrogen. These impurity elements in the core can be considered as additional scatterers for free electrons in metal. The effect of dilute impurity elements can be expressed by Matthiessen's rule:

$$\rho_{\text{Fe-alloy}}(V, T) = \rho_{\text{pure-Fe}}(V, T) + \sum_k \rho_{i,k} \cdot \chi_k \quad (4)$$

where $\rho_{\text{Fe-alloy}}$, $\rho_{\text{pure-Fe}}$, $\rho_{i,k}$ and χ_k are the electrical resistivities of iron alloy and pure iron, the impurity resistivity of element k , and the concentration of element k in iron, respectively. Matthiessen's rule indicates that element k contributes proportionally to its concentration χ_k independently of temperature. Note that this rule does not consider the effect of resistivity saturation, although the measured electrical resistivity of the Fe–Si alloy does show saturation phenomena^{40,41}. The high-temperature electrical resistivity, taking the resistivity saturation into account, is well described by a shunt resistor model (see equation (2) and ref. 23).

The impurity resistivity of Si in iron at high pressure has been examined using high-pressure experiments^{7,8}. According to Gomi *et al.*⁷, the impurity resistivity of Si $\rho_{\text{Si}}(V)$, in units of microOhm centimetres per atomic per cent, was formulated as:

$$\rho_{\text{Si}}(V) = F_1 \cdot \left(F_2 - \frac{V}{V_0} \right)^{F_3} \quad (5)$$

where V is the volume of ϵ iron at high pressure and V_0 is the volume of ϵ iron at 1 bar, respectively²¹. The fitting parameters in equation (5) were determined to be

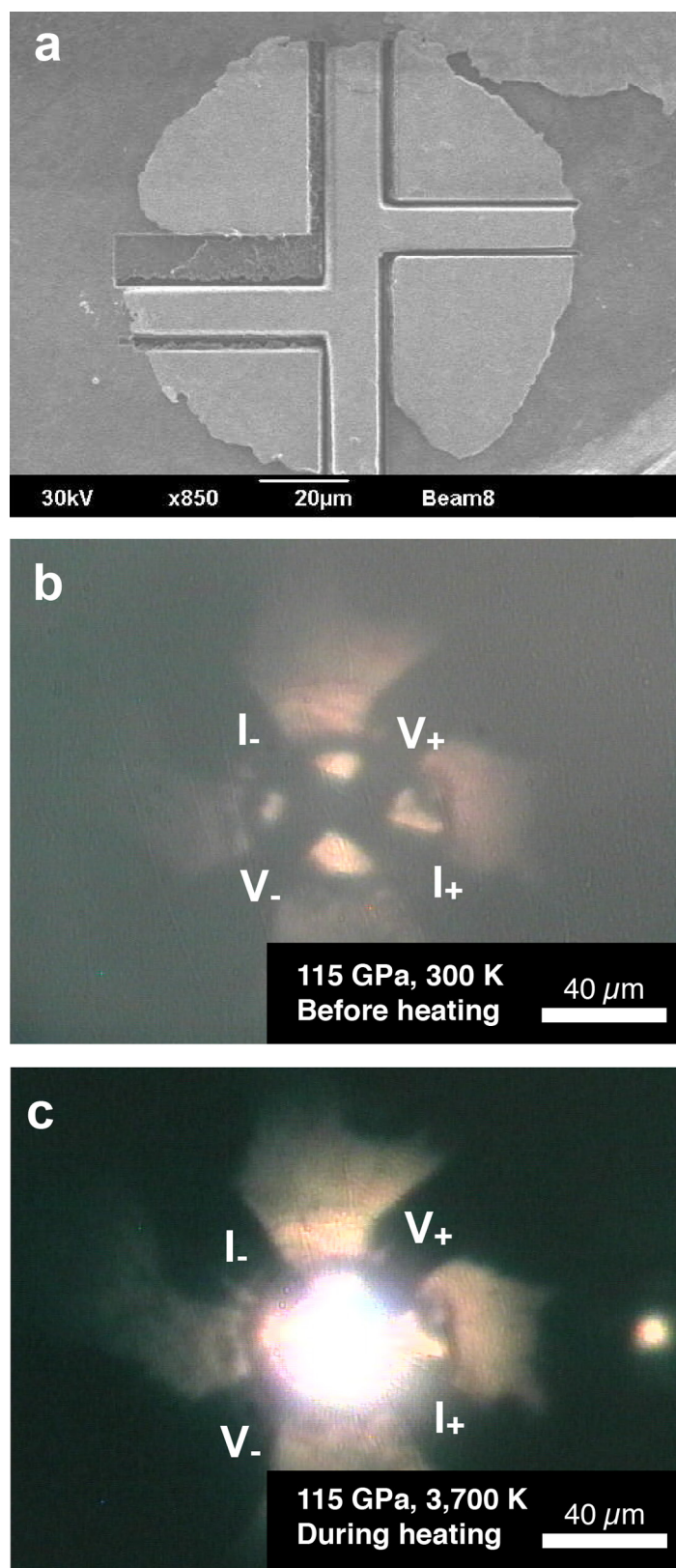
$F_1 = 3.77^{+2.52}_{-1.56} \mu\Omega \text{ cm/at\%}$, $F_2 = 1.48^{+0.11}_{-0.03}$, and $F_3 = -3^{+0.22}_{-0.25}$. In addition, Gomi and Hirose⁹ reported the volume-dependent impurity resistivity of Ni in ϵ iron at high pressures, $\rho_{\text{Ni}}(V)$, in units of microOhm centimetres per atomic per cent:

$$\rho_{\text{Ni}}(V) = F_4 \cdot \left(F_5 - \frac{V}{V_0} \right)^{F_6} \quad (6)$$

where $F_4 = 7.25^{+2.44}_{-1.08} \times 10^3 \mu\Omega \text{ cm/at\%}$, $F_5 = 3.51^{+0.13}_{-0.20}$ and $F_6 = -8.06^{+0.62}_{-1.19}$.

At 140 GPa, the impurity resistivities of Si and Ni are calculated to be $9.60^{+1.20}_{-2.63} \mu\Omega \text{ cm/at\%}$ and $1.97^{+0.27}_{-0.40} \mu\Omega \text{ cm/at\%}$, respectively, according to equations (5) and (6). Taking into account the resistivity saturation effect, we calculated the electrical resistivity of solid $\text{Fe}_{67.5}\text{Ni}_{10}\text{Si}_{22.5}$, the composition accounting for the outer core density²⁶, at 140 GPa. Since the saturation resistivity of solid $\text{Fe}_{67.5}\text{Ni}_{10}\text{Si}_{22.5}$ at 140 GPa is not known, we assume it to be $123^{+42}_{-28} \mu\Omega \text{ cm}$, the same as that for pure iron as determined in this study.

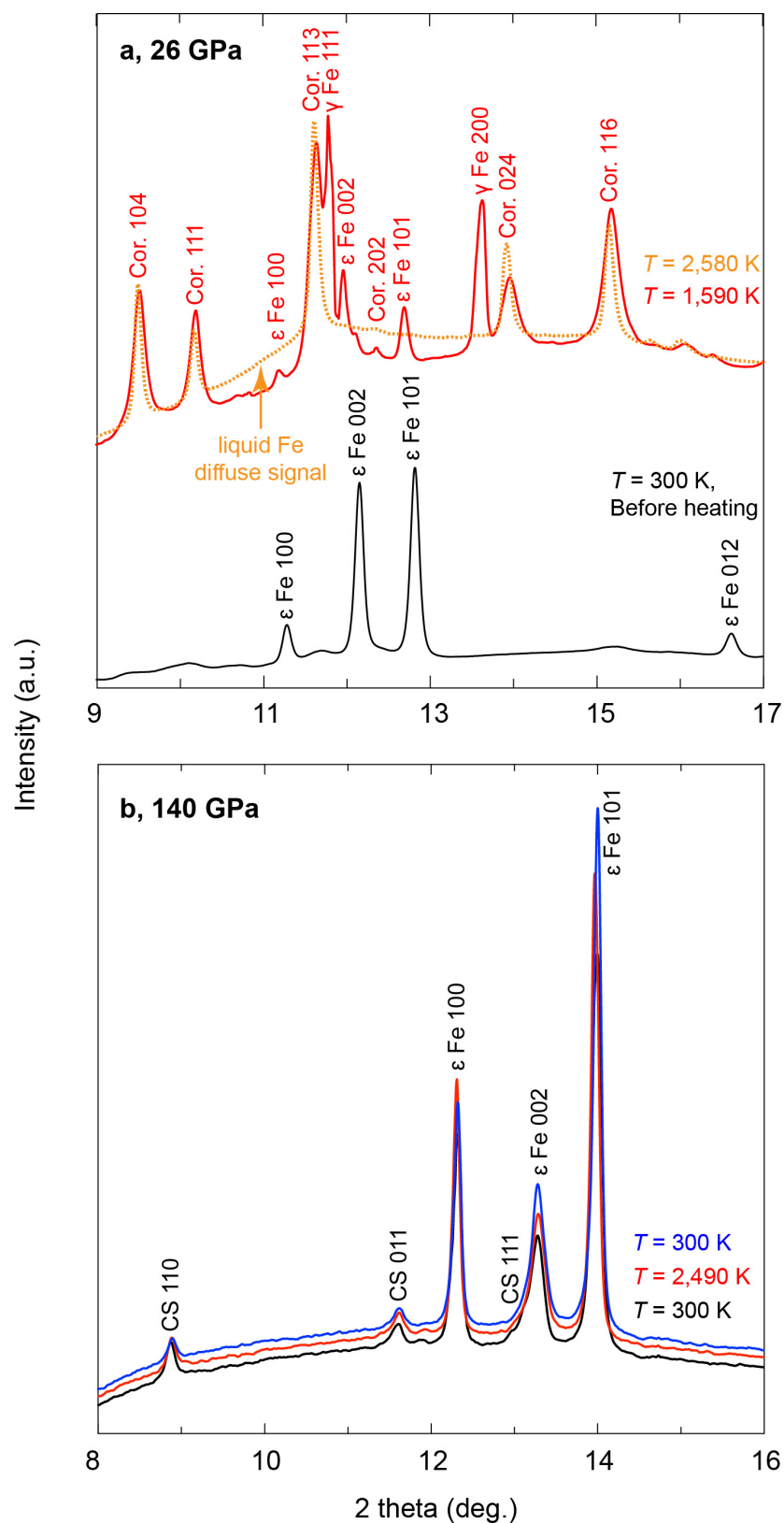
31. Akahama, Y. & Kawamura, H. Pressure calibration of diamond anvil Raman gauge to 310 GPa. *J. Appl. Phys.* **100**, 043516 (2006).
32. Ohishi, Y., Hirao, N., Sata, N., Hirose, K. & Takata, M. Highly intense monochromatic X-ray diffraction facility for high-pressure research at SPring-8. *High Press. Res.* **28**, 163–173 (2008).
33. Tsujino, N. *et al.* Equation of state of γ -Fe: reference density for planetary cores. *Earth Planet. Sci. Lett.* **375**, 244–253 (2013).
34. Jaccard, D., Holmes, A., Behr, G., Inada, Y. & Onuki, Y. Superconductivity of ϵ -Fe: complete resistive transition. *Phys. Lett. A* **299**, 282–286 (2002).
35. Touloukian, Y. S. in *Thermophysical Properties of Matter: Thermal Conductivity-Nonmetallic Solids* Ch. 2 (John Wiley and Sons, 1970).
36. Braginsky, L., Shklover, V., Hofmann, H. & Bowen, P. High-temperature thermal conductivity of porous Al_2O_3 nanostructures. *Phys. Rev. B* **70**, 134201 (2004).
37. Simpson, A. & Sticks, A. D. Study on thermal conductivity of boron nitride in hexagonal structure in atomistic scale by using non-equilibrium molecular dynamics technique. *J. Phys. C* **4**, 1710–1718 (1971).
38. Ohta, K. *et al.* Lattice thermal conductivity of MgSiO_3 perovskite and post-perovskite at the core–mantle boundary. *Earth Planet. Sci. Lett.* **349/350**, 109–115 (2012).
39. Imada, S. *et al.* Measurements of lattice thermal conductivity of MgO to core–mantle boundary pressures. *Geophys. Res. Lett.* **41**, 4542–4547 (2014).
40. Matasov, G. The electrical conductivity of iron–silicon alloys at high pressures and the Earth's core. PhD thesis. <https://e-reports-ext.llnl.gov/pdf/176480.pdf>, Lawrence Livermore Lab., Univ. California (1977).
41. Kiarasi, S. & Secco, R. Pressure-induced electrical resistivity saturation of Fe_{17}Si . *Phys. Status Solidi B* **252**, 2034–2042 (2015).
42. Boehler, R. Temperatures in the Earth's core from melting-point measurements of iron at high static pressures. *Nature* **363**, 534–536 (1993).
43. Anzellini, S., Dewaele, A., Mezouar, M., Loubeyre, P. & Morard, G. Melting of iron at Earth's inner core boundary based on fast X-ray diffraction. *Science* **340**, 464–466 (2013).
44. Ohno, H. Antiferromagnetism in hcp iron–ruthenium and hcp iron–osmium alloys. *J. Phys. Soc. Jpn.* **31**, 92–101 (1971).



Extended Data Figure 1 | Images of iron sample and sample configuration for electrical resistance measurements in a DAC.

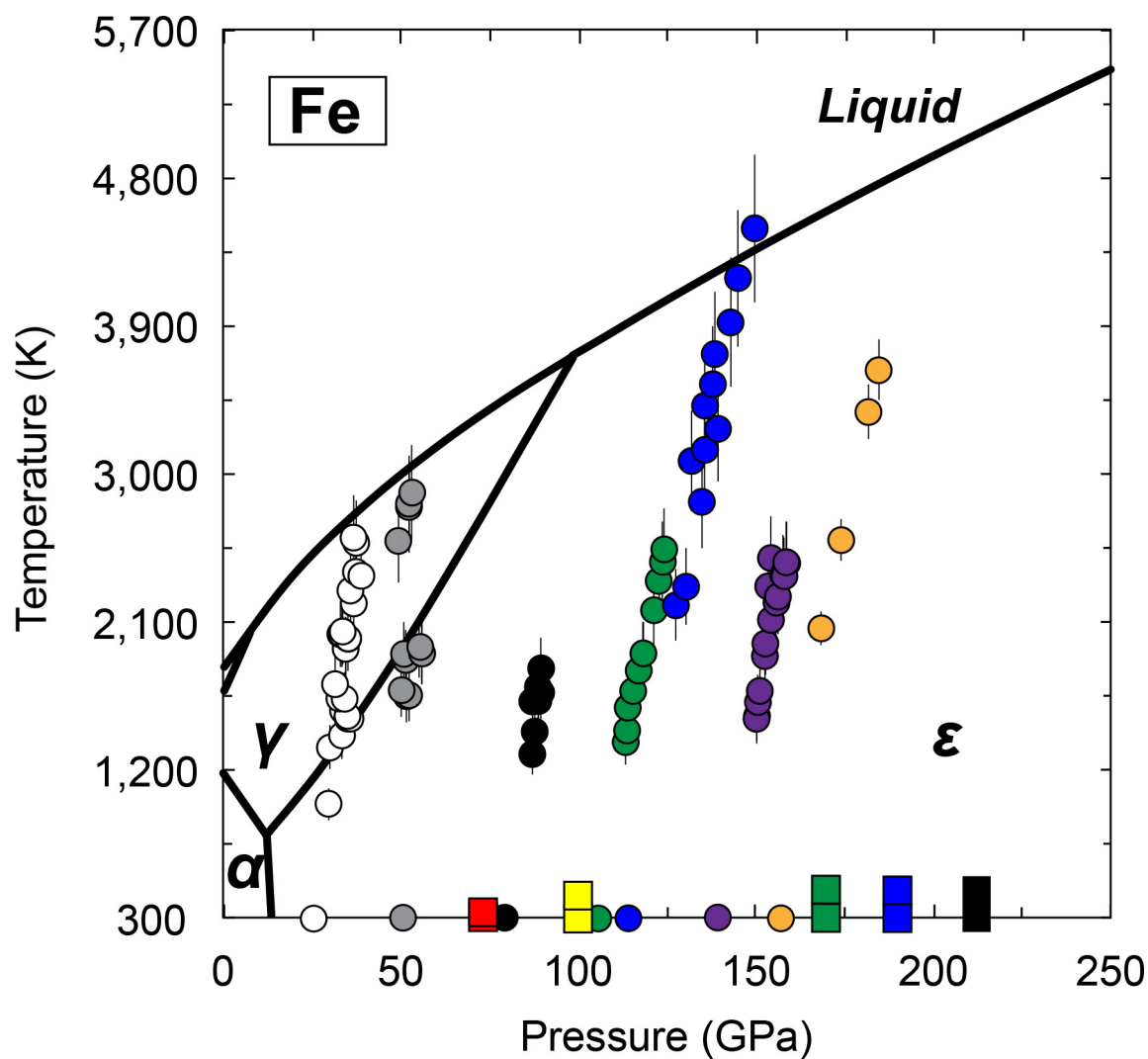
a, A composite of an iron sample and electrodes shaped by a focused ion beam. **b**, **c**, Photomicrographs of a sample chamber viewed through a diamond anvil at 115 GPa and 300 K (**b**) and 3,700 K (**c**). The four-probe method was used for electrical resistance measurements. At each set of

pressure and temperature conditions, we measured the voltage difference between two potential leads (V_+ and V_-) twice, when electric current passed through the sample from a positive current (I_+) lead to a negative current (I_-) lead and in the opposite direction. These two voltage values were averaged to eliminate thermal voltage, and the resistance is calculated from Ohm's law.



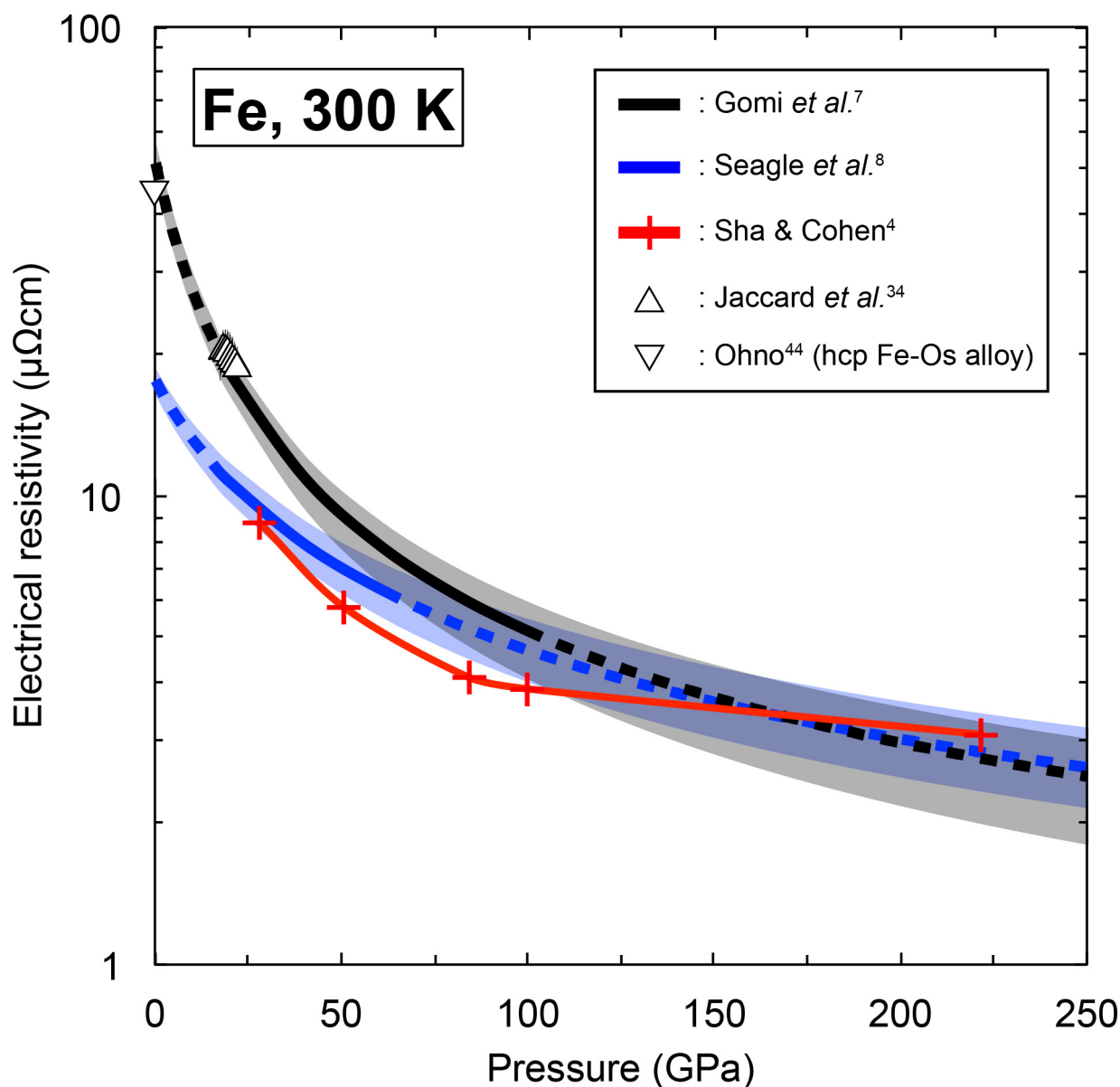
Extended Data Figure 2 | XRD patterns of iron samples at high pressures and temperatures. **a**, Data collected at 26 GPa (see Fig. 1a for resistivity measurement), showing the diffraction peaks of ϵ Fe, γ Fe, and Al_2O_3 pressure medium (Cor.). Liquid Fe is indicated by a diffuse signal at

$2\theta = 10^\circ$ to 14° . **b**, ϵ Fe at 140 GPa (Fig. 2e). Part of the SiO_2 glass pressure medium crystallized into a CaCl_2 -type phase (labelled as CS) when thermal annealing occurred at 80 GPa.



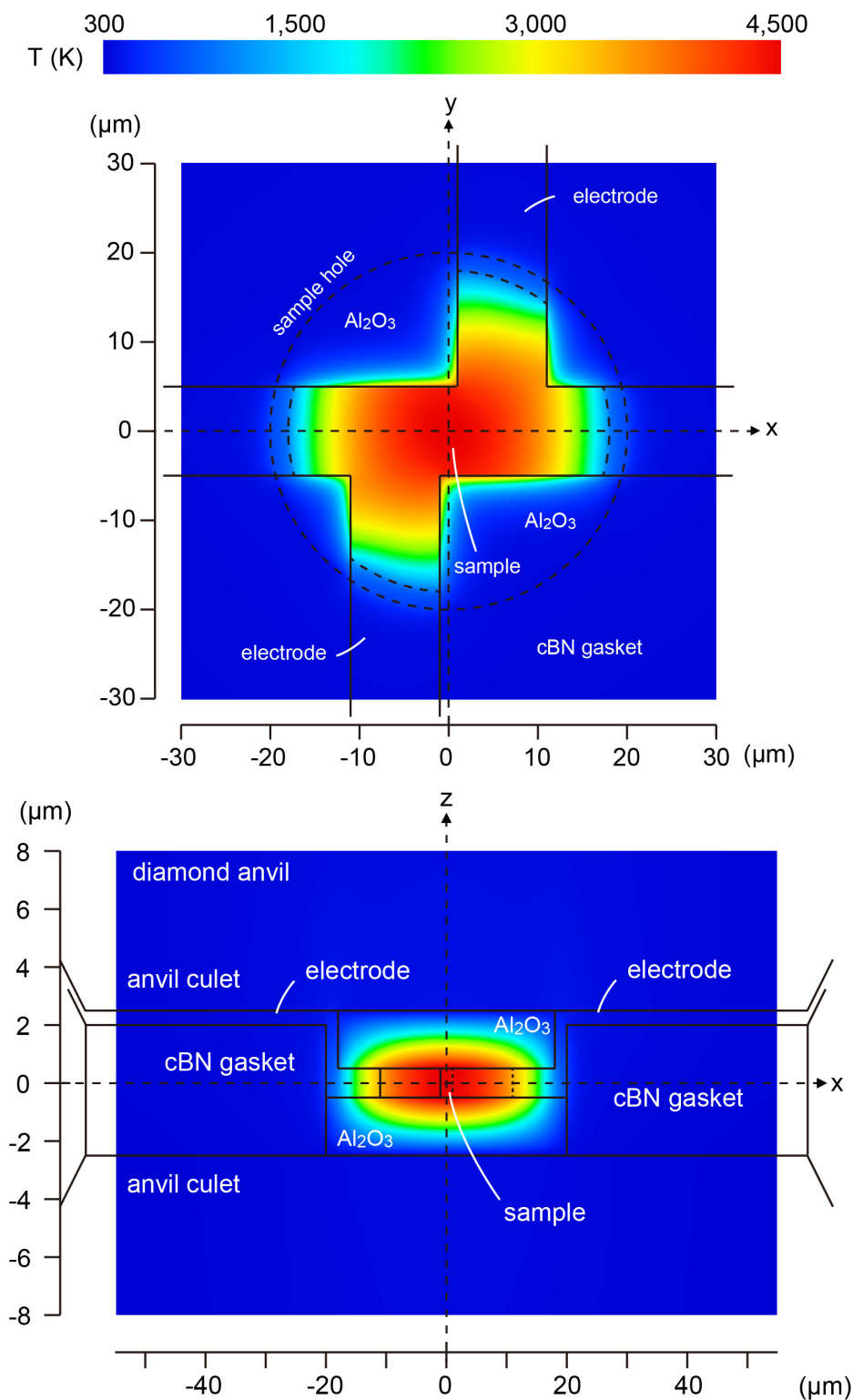
Extended Data Figure 3 | Pressure and temperature conditions of electrical resistivity measurements of iron. Circles show the conditions of laser-heated DAC experiments (Figs 1 and 2b–f), and squares indicate

those for measurements in a muffle furnace up to 450 K (Fig. 2a). Phase boundaries are from the literature^{42,43}. The symbol colours in this figure correspond to those in Figs 1 and 2. Error bars, 1 σ .



Extended Data Figure 4 | Electrical resistivity of ϵ iron at high pressure and 300 K. Bold black and blue curves show the results of previous DAC experiments by Gomi *et al.*⁷ and Seagle *et al.*⁸, respectively, with uncertainties shown by bands. The red curve connecting crosses is from

theoretical calculations⁴. All these results are consistent with each other above ~ 80 GPa. For comparison, the resistivity of ϵ iron at 1 bar deduced from the measurements of hexagonal close-packed (hcp) Fe-Os alloy⁴⁴ and at low pressures in a DAC³⁴ are also shown.



Culet size : 120 μm

115 GPa, 4,500 K

Extended Data Figure 5 | Temperature maps of the iron sample and electrodes in a laser-heated DAC at 115 GPa and 4,500 K. Al_2O_3 is the pressure medium. The top panel shows the temperature map viewed along the compression axis. The bottom panel shows the temperature map

of the cross-section of a sample chamber and a gasket. The maximum temperature difference in the area for resistance measurement is 200 K, smaller than the uncertainty in the temperature determination.

Extended Data Table 1 | Saturation resistivity ρ_{sat} of ε iron at high pressures

P (GPa)*	$(V/V_0)^{1/3}$	ρ_{sat} ($\mu\Omega\text{cm}$)	Reference
0	1.000	168	ref. 24 [†]
80	0.919	142(+32/-26)	This study
106	0.904	136(+53/-27)	
114	0.901	124(+35/-26)	
140	0.890	123(+42/-28)	
157	0.883	122(+23/-29)	

*Calculated from the equation of state of ε iron²¹.[†]Ref. 24 determined the ρ_{sat} of iron to be $168\mu\Omega\text{ cm}$.

Direct measurement of thermal conductivity in solid iron at planetary core conditions

Zuzana Konôpková^{1†}, R. Stewart McWilliams², Natalia Gómez-Pérez^{2,3} & Alexander F. Goncharov^{4,5}

The conduction of heat through minerals and melts at extreme pressures and temperatures is of central importance to the evolution and dynamics of planets. In the cooling Earth's core, the thermal conductivity of iron alloys defines the adiabatic heat flux and therefore the thermal and compositional energy available to support the production of Earth's magnetic field via dynamo action^{1–3}. Attempts to describe thermal transport in Earth's core have been problematic, with predictions of high thermal conductivity^{4–7} at odds with traditional geophysical models and direct evidence for a primordial magnetic field in the rock record^{8–10}. Measurements of core heat transport are needed to resolve this difference. Here we present direct measurements of the thermal conductivity of solid iron at pressure and temperature conditions relevant to the cores of Mercury-sized to Earth-sized planets, using a dynamically laser-heated diamond-anvil cell^{11,12}. Our measurements place the thermal conductivity of Earth's core near the low end of previous estimates, at 18–44 watts per metre per kelvin. The result is in agreement with palaeomagnetic measurements¹⁰ indicating that Earth's geodynamo has persisted since the beginning of Earth's history, and allows for a solid inner core as old as the dynamo.

The thermal evolution of Earth's core and the energetics of the geomagnetic field are highly sensitive^{3,8,9} to the thermal conductivity of core materials at the high pressures (P) and high temperatures (T) of the core. A wide range of values for the thermal conductivity of iron (Fe) and its alloys at core conditions have been predicted using materials theory^{2,4,6,7,13} and high-pressure measurements of electrical conductivity^{5,14–16}. To

predict thermal conductivity, the Wiedemann–Franz–Lorenz law:

$$k = LT\sigma \quad (1)$$

has almost universally been employed, where k and σ are the thermal and electrical conductivities and L is the Lorenz number. The Lorenz number—traditionally an empirically determined quantity¹⁷—has been calculated theoretically^{6,7} but not measured for Fe or its alloys at high pressure and temperature conditions.

For low estimates of thermal conductivity², near $k = 30 \text{ W m}^{-1} \text{ K}^{-1}$, the geodynamo may be sustained during the whole life of the planet, and convection of the core is readily attained in thermal (in absence of an inner core) or thermochemical scenarios⁹. On the other hand, a recent estimate⁶ near $k = 130 \text{ W m}^{-1} \text{ K}^{-1}$ implies a young inner core (that is, less than 1.3 billion years old), and only thermal convection driving the dynamo at earlier times³. However, a paradox arises⁸ when evidence of an ancient magnetic field^{3,10} must be reconciled with the high energy fluxes needed to drive thermal convection in a high conductivity, fully fluid core. The large core–mantle boundary heat flux (Q_{CMB}) and high internal temperatures for the early Earth in this case (implying a molten lower mantle and possibly a stably stratified core) are difficult to explain given current mantle evolution models and low present-day Q_{CMB} (ref. 3). Re-evaluating the history and energy balances of Earth's core and mantle in this context, it is necessary to have certainty on the validity of reported values of k (ref. 8). Thus, there is a pressing need for direct thermal conductivity measurements of core materials at conditions relevant to Earth's core.

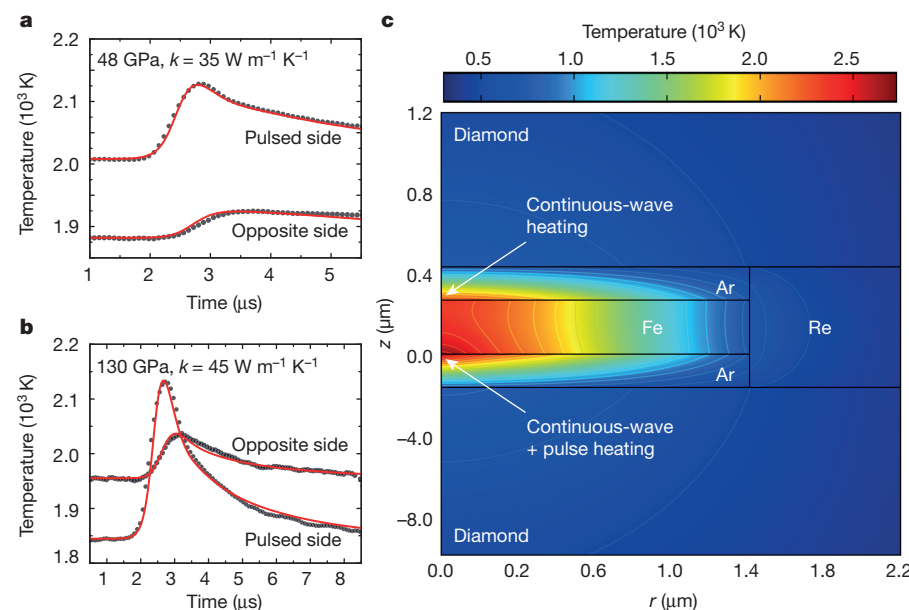


Figure 1 | Temperature of Fe foils during flash heating at high initial temperature and pressure. **a, b,** Plots of the measured temperature histories (grey) on the pulsed and opposite sides of the foil together with finite-element models (red) for best-fit thermal conductivity k of Fe, at two pressures: **a**, $P = 48 \text{ GPa}$; **b**, $P = 130 \text{ GPa}$. **c**, Instantaneous temperature map of the modelled sample area at initiation of flash heating at 112 GPa , as a function of radial (r) and axial (z) position. Contour lines are isotherms.

¹DESY Photon Science, Notkestrasse 85, DE-22607 Hamburg, Germany. ²School of Physics and Astronomy and Centre for Science at Extreme Conditions, University of Edinburgh, Peter Guthrie Tait Road, Edinburgh EH9 3FD, UK. ³Departamento de Geociencias, Universidad de Los Andes, Bogotá, Colombia. ⁴Key Laboratory of Materials Physics, Institute of Solid State Physics, Chinese Academy of Sciences, 350 Shushanghu Road, Hefei, Anhui 230031, China. ⁵Geophysical Laboratory, Carnegie Institution of Washington, 5251 Broad Branch Road NW, Washington DC 20015, USA. [†]Present address: European XFEL GmbH, Notkestrasse 85, DE-22607 Hamburg, Germany.

Although the technical capability of reaching planetary core conditions in the laboratory has long been available using the laser-heated diamond-anvil cell (DAC), measurements sensitive to transport properties have been scarce. Thermal transport measurements have been especially challenging. To overcome this limitation, we dynamically measured temperature in the laser-heated DAC^{11,12} to study the propagation of heat pulses across Fe foils contained at high initial pressure (35–130 GPa) and temperature (1,600–3,000 K) (Fig. 1). Fitting of the temporally and spatially resolved temperature fluctuations with heat conduction models provides a strong constraint on the thermal transport (Methods and Extended Data Figs 2–6).

The experiments performed below ~50 GPa probe Fe in the stability field of face-centred cubic γ Fe (Fig. 2)^{18–22}. At conditions close to those at the centre of Mercury's core²³ (~40 GPa and 2,200–2,500 K), thermal conductivity is $35 \pm 10 \text{ W m}^{-1} \text{ K}^{-1}$. This is similar to the ambient pressure values in γ Fe ($k = 30 \pm 3 \text{ W m}^{-1} \text{ K}^{-1}$)²⁴, suggesting that k is not strongly dependent on pressure at Mercury's core conditions. This result is similar to earlier expectations for the thermal conductivity of Mercury's core²⁵ of $\sim 40 \text{ W m}^{-1} \text{ K}^{-1}$, but is at odds with more recent estimates²¹. At pressures in the range 50–80 GPa, the sample is usually pre-heated in the hexagonal close-packed ϵ Fe phase but may undergo partial transformation to the γ phase during the thermal pulse. Thermal conductivity values found at these conditions are considered biased towards the ϵ phase, and are in general agreement with earlier DAC measurements on ϵ Fe (ref. 26). The highest-pressure data, 88–130 GPa at 1,600–3,500 K, are unambiguously in the region of ϵ Fe and are closest to the conditions at Earth's core–mantle boundary^{1,6}: 136 GPa and 3,800–4,800 K. A large number of measurements (>20) at 112 GPa show k to decrease with temperature at these conditions (Fig. 3), as expected from combining electrical conductivity data under static and shock wave compression¹⁴.

To model the temperature dependence of thermal conductivity in ϵ Fe, we fitted the data at 112 GPa to:

$$k = aT + \frac{b}{\sqrt{T}} \quad (2)$$

This form ensures a realistic behaviour of both thermal conductivity and electrical resistivity ($1/\sigma$) that is consistent with previous high-temperature resistivity data^{5,14,21} (see Methods and Extended Data Fig. 1). The model fit at 112 GPa (Fig. 3) also includes resistivity data at room temperature^{5,14} extrapolated to 112 GPa and shock wave resistivity data¹⁵ interpolated to 112 GPa. These data were converted to thermal conductivity using an empirical Lorenz number of $(1.9 \pm 0.4) \times 10^{-8} \text{ W } \Omega \text{ K}^{-2}$ (see Methods). The fit of equation (2) yields $b \approx 1,972 \text{ W m}^{-1} \text{ K}^{-1/2}$ and $a \approx 0$. The error in model thermal conductivities is ~20% (one standard deviation).

To assess the pressure variation of k in ϵ Fe, we used a physical model for the variation of electronic thermal conductivity with pressure (see Methods) in terms of isothermal bulk modulus (K_T) and Grüneisen parameter (γ):

$$\frac{1}{k} \frac{\partial k}{\partial P} = \frac{2\gamma - 1/3}{K_T} \quad (3)$$

The Grüneisen parameter and bulk modulus at core conditions are evaluated using the thermal equation of state of Fe (ref. 27) (see Methods). The model represents our data well to 130 GPa (Fig. 2), and predicts somewhat larger values of k at Earth's outer core conditions (Fig. 3). Accounting for the uncertainty in outer core temperature^{1,6}, k for pure Fe varies from $33 \pm 7 \text{ W m}^{-1} \text{ K}^{-1}$ at core–mantle boundary conditions ($T = 3,800\text{--}4,800 \text{ K}$, $P = 136 \text{ GPa}$) to $46 \pm 9 \text{ W m}^{-1} \text{ K}^{-1}$ at inner-core boundary conditions ($T = 5,600\text{--}6,500 \text{ K}$, $P = 330 \text{ GPa}$).

The conductivity of molten Fe, which is relevant to the outer core, is generally taken to be similar to that of solid Fe near melting^{13,21,28}. The addition of light-element impurities is expected to reduce conductivity by 10%–40% (refs 7 and 13). Thus, the thermal conductivity

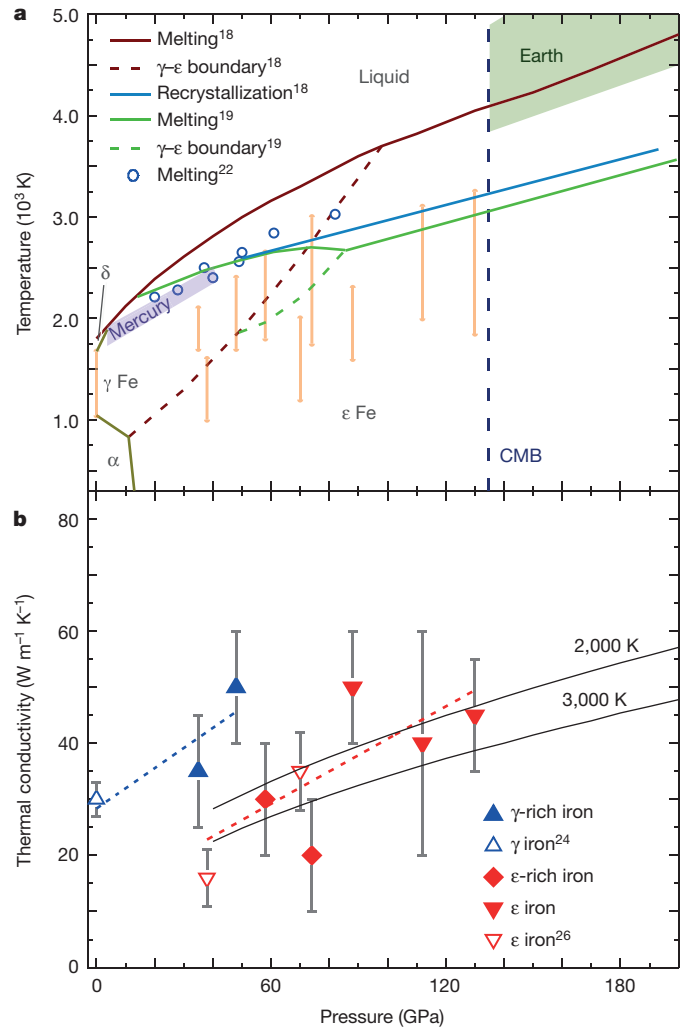


Figure 2 | Thermal conductivity of Fe at high pressure and temperature.

a, Phase diagram^{18–20,22} of Fe with conditions of the thermal conductivity measurements (orange) falling in the domain of the γ and ϵ phases. The shaded areas depict conditions of Earth's core^{1,6} and Mercury's core²³, with the vertical dashed line marking the pressure at Earth's core–mantle boundary (CMB). **b**, Thermal conductivity results from this study are shown as solid symbols: in the domain of γ Fe (upward triangles) the γ and ϵ phases most probably co-exist¹⁸; for samples typically pre-heated to below the γ – ϵ boundary and then crossing it briefly during thermal pulses (diamonds), the phase is considered to be mostly ϵ Fe; at higher pressure (downward triangles) samples are pure ϵ Fe at all conditions¹⁸ (see Methods). Prior direct thermal conductivity measurements on the γ phase²⁴ and the ϵ phase²⁶ are shown as open symbols. The dashed lines are linear fits to the results from the γ and ϵ domains, whereas solid lines are model values (see equations (2) and (3)). Error bars include uncertainty (one standard deviation) and range of measurements.

for Earth's liquid outer core is between $25 \pm 7 \text{ W m}^{-1} \text{ K}^{-1}$ at the core–mantle boundary and $35 \pm 10 \text{ W m}^{-1} \text{ K}^{-1}$ at the inner-core boundary. Refining estimates for liquid core composition can further reduce this uncertainty. The corresponding electrical resistivity of the outer core is $3.7 \pm 1.5 \mu\Omega \text{ m}$.

Our thermal conductivities for pure Fe at core conditions compare well with predictions based on resistivity measurements at high pressure¹⁴ including shock wave results ($52 \pm 11 \text{ W m}^{-1} \text{ K}^{-1}$) or Stacey's law of constant resistivity at melting² ($48 \pm 10 \text{ W m}^{-1} \text{ K}^{-1}$), where the empirical value of L has been applied. Such predictions are sensitive to the assumptions used, however, and much larger values are found using slightly different approaches^{5,13,14}, emphasizing the need for direct constraints from high-pressure, high-temperature data. Calculations^{6,7} finding $k = 120\text{--}160 \text{ W m}^{-1} \text{ K}^{-1}$ at core–mantle boundary conditions

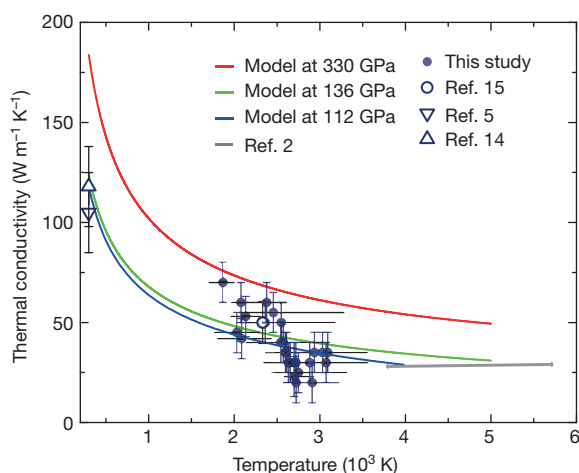


Figure 3 | Thermal conductivity of Fe versus temperature. Solid circles indicate results from this study at 112 GPa, with horizontal bars indicating the range of temperatures observed in each experiment, and vertical bars the uncertainty in k (one standard deviation). Estimates based on prior electrical resistivity measurements^{5,14,15} are shown as open symbols, with bars indicating uncertainty from the empirical determination of L . The thermal conductivity model for 112 GPa, 136 GPa (core–mantle boundary), and 330 GPa (inner–core boundary) are blue, green and red lines, respectively (see equations (2) and (3)). For comparison, the prediction of ref. 2 for core alloy at outer-core conditions is the grey line.

and $k = 205\text{--}250\text{ W m}^{-1}\text{ K}^{-1}$ at inner-core boundary conditions are 5.6 ± 1.8 and 6.5 ± 1.7 times larger than our values, respectively.

During an early stage of Earth history before the formation of the inner core, the presence of the geodynamo requires a core–mantle boundary heat flux (Q_{CMB}) greater than the conductive heat flux in the core. The heat flux requirements for such a convective early core are moderate for the values of k found in this study, similar to that of ref. 9: Q_{CMB} must exceed a threshold of $3.8 \pm 1.6\text{ TW}$ (for k of $31 \pm 13\text{ W m}^{-1}\text{ K}^{-1}$) for Earth's magnetic field to be sustained, assuming negligible radiogenic heating. Later in the planet's history, after a solid inner core has formed, the core–mantle heat flux necessary to sustain a dynamo may be smaller, given that convection can be driven both compositionally and thermally. Estimates³ for the current Q_{CMB} ($12 \pm 5\text{ TW}$) far exceed this threshold, so for a nominal scenario of Q_{CMB} declining or constant with time^{3,9} magnetic activity is expected throughout Earth history, and would probably only have been absent when internal dynamics differed substantially from those of the present, for example in periods lacking plate tectonics²⁹. Similarly, evidence of non-zero palaeomagnetic field places a hard constraint on the corresponding heat flux of $Q_{\text{CMB}} > 2.2\text{ TW}$ before inner-core nucleation.

However, the inner core can be older for lower core thermal conductivities⁵, and within the uncertainty due to the light-element content of the core, the inner core can be as old as the earliest recorded terrestrial magnetic field¹⁰, that is, up to 4.2 billion years old. Thus, within our direct experimental constraints, there is no requirement that Earth's geodynamo ever existed in the absence of an inner core. Indeed, the planet's dynamo and its solid inner core may have co-existed since soon after the formation of Earth. Greater knowledge of the light-element content of the core and its effect on thermal conductivity is essential to understand the earliest period of Earth's core evolution.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 30 September 2015; accepted 11 April 2016.

- Labrosse, S. Thermal and magnetic evolution of the Earth's core. *Phys. Earth Planet. Inter.* **140**, 127–143 (2003).
- Stacey, F. D. & Loper, D. E. A revised estimate of the conductivity of iron alloy at high pressure and implications for the core energy balance. *Phys. Earth Planet. Inter.* **161**, 13–18 (2007).

- Nimmo, F. in *Treatise on Geophysics* 2nd edn (ed. Schubert, G.) 27–55, 201–219 (Elsevier, 2015).
- Sha, X. & Cohen, R. First-principles studies of electrical resistivity of iron under pressure. *J. Phys. Condens. Matter* **23**, 075401 (2011).
- Gomi, H. et al. The high conductivity of iron and thermal evolution of the Earth's core. *Phys. Earth Planet. Inter.* **224**, 88–103 (2013).
- Pozzo, M., Davies, C., Gubbins, D. & Alfe, D. Thermal and electrical conductivity of iron at Earth's core conditions. *Nature* **485**, 355–358 (2012).
- de Koker, N., Steinle-Neumann, G. & Vlcek, V. Electrical resistivity and thermal conductivity of liquid Fe alloys at high P and T, and heat flux in Earth's core. *Proc. Natl Acad. Sci. USA* **109**, 4070–4073 (2012).
- Olson, P. The new core paradox. *Science* **342**, 431–432 (2013).
- Nimmo, F. in *Treatise on Geophysics* (ed. Schubert, G.) 31–65, 217–241 (Elsevier, 2007).
- Tarduno, J. A., Cottrell, R. D., Davis, W. J., Nimmo, F. & Bono, R. K. A Hadean to Paleoproterozoic geodynamo recorded by single zircon crystals. *Science* **349**, 521–524 (2015).
- McWilliams, R. S., Konôpková, Z. & Goncharov, A. F. A flash heating method for measuring thermal conductivity at high pressure and temperature: application to Pt. *Phys. Earth Planet. Inter.* **247**, 17–26 (2015).
- McWilliams, R. S., Dalton, D. A., Konôpková, Z., Mahmood, M. F. & Goncharov, A. F. Opacity and conductivity measurements in noble gases at conditions of planetary and stellar interiors. *Proc. Natl Acad. Sci. USA* **112**, 7925–7930 (2015).
- Stacey, F. D. & Anderson, O. L. Electrical and thermal conductivities of Fe–Ni–Si alloy under core conditions. *Phys. Earth Planet. Inter.* **124**, 153–162 (2001).
- Seagle, C. T., Cottrell, E., Fei, Y. W., Hummer, D. R. & Prakapenka, V. B. Electrical and thermal transport properties of iron and iron–silicon alloy at high pressure. *Geophys. Res. Lett.* **40**, 5377–5381 (2013).
- Bi, Y., Tan, H. & Jing, F. Electrical conductivity of iron under shock compression up to 200 GPa. *J. Phys. Condens. Matter* **14**, 10849 (2002).
- Keeler, R. N. & Royce, E. B. in *Physics of High Energy Density* Vol. 48 (eds Caldirola, P. & Knoepfel, H.) 106–125 (Academic Press, 1971).
- Franz, R. & Wiedemann, G. Ueber die Wärme-Leitungsfähigkeit der Metalle. *Ann. Phys.* **165**, 497–531 (1853).
- Anzellini, S., Dewaele, A., Mezouar, M., Loubeyre, P. & Morard, G. Melting of iron at Earth's inner core boundary based on fast X-ray diffraction. *Science* **340**, 464–466 (2013).
- Boehler, R. Temperatures in the Earth's core from melting-point measurements of iron at high static pressures. *Nature* **363**, 534–536 (1993).
- Komabayashi, T., Fei, Y., Meng, Y. & Prakapenka, V. In-situ X-ray diffraction measurements of the γ - ϵ transition boundary of iron in an internally-heated diamond anvil cell. *Earth Planet. Sci. Lett.* **282**, 252–257 (2009).
- Deng, L., Seagle, C., Fei, Y. & Shahar, A. High pressure and temperature electrical resistivity of iron and implications for planetary cores. *Geophys. Res. Lett.* **40**, 33–37 (2013).
- Jackson, J. M. et al. Melting of compressed iron by monitoring atomic dynamics. *Earth Planet. Sci. Lett.* **362**, 143–150 (2013).
- Rivoldini, A., Van Hoolst, T. & Verhoeven, O. The interior structure of Mercury and its core sulfur content. *Icarus* **201**, 12–30 (2009).
- Ho, C. Y., Powell, R. W. & Liley, P. E. Thermal conductivity of the elements. *J. Phys. Chem. Ref. Data* **1**, 279–422 (1972).
- Hauck, S. A., Dombard, A. J., Phillips, R. J. & Solomon, S. C. Internal and tectonic evolution of Mercury. *Earth Planet. Sci. Lett.* **222**, 713–728 (2004).
- Konôpková, Z., Lazor, P., Goncharov, A. F. & Struzhkin, V. V. Thermal conductivity of hcp iron at high pressure and temperature. *High Press. Res.* **31**, 228–236 (2011).
- Dubrovinsky, L. S., Saxena, S. K., Tutti, F., Rekhi, S. & LeBehan, T. In situ X-ray study of thermal expansion and phase transition of iron at multimegabar pressure. *Phys. Rev. Lett.* **84**, 1720–1723 (2000).
- Secco, R. A. & Schloessin, H. H. The electrical resistivity of solid and liquid Fe at pressures up to 7 GPa. *J. Geophys. Res. Solid Earth* **94**, 5887–5894 (1989).
- Nimmo, F. & Stevenson, D. J. Influence of early plate tectonics on the thermal evolution and magnetic field of Mars. *J. Geophys. Res. Planets* **105**, 11969–11979 (2000).

Acknowledgements We acknowledge experimental assistance from H. Marquardt. This work was supported by the NSF (grant numbers DMR-1039807, EAR-1015239, EAR-1520648 and EAR/IF-1128867), the Army Research Office (grant 56122-CH-H), the Carnegie Institution of Washington, the National Natural Science Foundation of China (grant number 21473211), the Chinese Academy of Science (grant number YZ201524), the University of Edinburgh, and the British Council Researcher Links Programme. Portions of this research were carried out at the light source Petra III at DESY, a member of the Helmholtz Association (HGF).

Author Contributions Z.K., R.S.M. and A.F.G. designed and conducted experiments. R.S.M. reduced raw data. Z.K. and N.G.P. performed finite-element modelling. N.G.P. performed error analysis and geophysical calculations. All authors wrote the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to Z.K. (zuzana.konopkova@xfel.eu), R.S.M. (rs.mcwilliams@ed.ac.uk), N.G.P. (ngomez@uniandes.edu.co) or A.F.G. (alex@issp.ac.cn).

METHODS

Briefly, a high-purity Fe foil (99.99%, GoodFellow) placed between two anvils of the DAC and separated from the anvils by layers of insulating material (NaCl or Ar) was preheated to a desired stable temperature using double-sided continuous-wave infrared laser heating, and then pulse-heated on one side with an additional infrared laser to create a thermal disturbance¹¹. The evolution of this disturbance was characterized by nanosecond-resolved radiative temperature measurements using a streak camera coupled to a grating spectrograph that records the thermal incandescent history from both sides of the foil. The phase shift and the reduction in amplitude of the temperature disturbance as it propagates across the foil are thus measured¹¹. At a given pressure, a series of data sets were collected using different continuous-wave and pulse laser powers. Temperatures studied ranged from ~1,600 K, the lowest detectable temperature, to 4,000 K at the maximum, whereas temperature disturbances were typically a few hundred kelvin in amplitude.

The temperature evolution was fitted to time-dependent finite-element models of the laser-heated DAC^{11,26,30} to determine the thermal conductivity of Fe samples. For the finite-element modelling we employed experimentally determined geometrical parameters and thermochemical parameters determined from known equations of state. The thermal conductivity of the sample, together with the thermal conductivity of the pressure medium and heating power, were adjusted until the best match of modelled and experimental temperature was achieved (Fig. 1a). The analysis was rigorously tested for sensitivity to input parameters (Extended Data Figs 2, 3 and 6). Total uncertainty and error bars (Fig. 2) were determined from the fitting uncertainty (Extended Data Figs 2 and 5), the scatter across different data sets (for example, Fig. 3), and uncertainty in input parameters (Extended Data Figs 3 and 6). We find the measurements to be sufficiently sensitive to the thermal conductivity of the sample foil to provide a major constraint on Fe conductivity at core conditions.

The experiment duration (less than 10 s per temperature history collection) was kept as short as possible to avoid sample damage and minimize the heating of optics and DAC that could cause instabilities during long laser-heating runs. Foil initial thickness ($4.01 \pm 0.02 \mu\text{m}$) and *in situ* thickness (Extended Data Table 1) were measured using white-light interferometry of the DAC cavity, and the index of refraction data for the media under pressure^{31–33}; these measurements also determined the sample-to-diamond-culet distances, which are important parameters in finite-element calculations. Foil thickness changes measured under compression were consistent with those derived from the known compressibility of Fe³⁴. For the NaCl medium, insulation plates were formed and placed on the culets, and foils were placed between them; in the case of Ar, the foil was suspended on a recess in the gasket (Re).

A sample of platinum, which has well defined thermal conductivity behaviour¹¹ at high pressure and temperature, was available as a control in some experiments at low pressures where the DAC cavities were sufficiently large in diameter ($P \leq 55 \text{ GPa}$) to accommodate a second foil. The Pt foil had the same thickness as the Fe foil, and was positioned on the plane of the Fe foil in the cavity; for such foil pairs, sample and insulation thicknesses, cell geometry, pressure, medium, heating configuration, and detection system were identical, allowing a direct relative comparison between the thermal transport behaviour of the two materials. Heat wave propagation across the Pt was much faster than for Fe (for example, 240 ns for the half-rise time, compared to 565 ns in Fe at 48 GPa; see Extended Data Fig. 4), corresponding to a lower thermal diffusivity for Fe. Fe samples were also observed to sustain larger axial temperature gradients than the Pt samples, manifested in a greater difference between peak amplitudes on either side of the foil. These observations affirm that at the studied conditions, the thermal conductivity of Pt ($160 \pm 40 \text{ W m}^{-1} \text{ K}^{-1}$; ref. 11) is substantially greater than that of Fe.

The Lorenz number for ϵ Fe was determined by comparing shock wave electrical resistivity¹⁵ and the present thermal conductivity data at comparable pressure and temperature (Fig. 3). The result is $22 \pm 16\%$ lower than the value for a free-electron metal³⁵ ($L = 2.44 \times 10^{-8} \text{ W } \Omega \text{ K}^{-2}$), consistent with theoretical calculations⁷, which predict a Lorenz number reduced from the ideal by up to 17%.

Experimental details. To generate thermal perturbations at high initial pressure and temperature, we combined double-sided continuous and single-sided pulsed laser heating of the DAC sample¹¹. The initial temperature was reached by balancing laser power to either side of the sample until temperatures agreed to within ~100 K, and then pulsed heating was used to create a small perturbation in temperature which propagated across the sample. Our approach is similar to that used in traditional flash heating measurements of thermal diffusivity³⁶, modified for a specimen under pressure in a DAC¹¹. The reduction in amplitude and phase shifting of the heat pulse with distance is an essentially one-dimensional phenomenon^{11,36}, whereas two-dimensional effects have a secondary, but non-negligible, impact accounted for via finite-element modelling.

Precise temperature determination during pulse laser heating was made with a streak camera detecting system coupled to a spectrometer, capable of detecting

thermal emission in a time-resolved manner in a spectrogram. Spectrograms ($3\text{--}10 \mu\text{s}$) were synchronized to the heating pulses to follow the sample's temperature response on both sides. Thermal emission was fitted to a greybody Planck function assuming constant emissivity during the heat cycle¹¹, a reasonable approximation since thermal perturbations are small. The time resolution of the temperature measurements was 26 ns ($3\text{--}\mu\text{s}$ sweep) to 82 ns ($10\text{--}\mu\text{s}$ sweep). Spectrograms were integrated over 10^2 to 10^4 perturbation cycles, at a rate of 1 kHz and total integration times of 0.1–10 s, the total integration time depending on temperature. Emission was calibrated to a tungsten ribbon lamp of known radiance. Temperatures were detected only above ~1,600 K owing to a lack of signal at lower temperatures. Experiments were limited at high temperatures owing to visible foil deformation in the melting regime of the sample and pressure medium¹¹.

Pressure was measured by the ruby fluorescence technique at room temperature. Thermal pressures produced during laser heating are positive but small (of the order of a few gigapascals) in sample configurations similar to those used here¹⁸ and do not significantly affect our results.

At pressures and temperatures in the stability field of γ Fe, face-centred cubic γ Fe and hexagonal close-packed ϵ Fe are commonly observed to coexist in experiments¹⁸. Consequently, our data at these conditions may probe a mixed state of γ Fe and ϵ Fe with a variable γ Fe composition (Fig. 2). In contrast, at higher pressures, ϵ Fe is typically the only observed solid phase at all temperatures^{18,37}, so our data in this regime directly probe pure ϵ Fe. To test these expectations, we have also performed *in situ* X-ray diffraction measurements on laser-heated Fe samples prepared in a manner identical to that used in this study (with NaCl media), at the P02.2 beamline (ECB) of PETRA III in Hamburg. Using comparable timescales of heating, we confirm that a mixed phase should be present in the lower-pressure experiments reported in this study, but not at higher pressures.

To prevent the uptake of impurities in our initially high-purity Fe foils, pressure-medium materials (NaCl, Ar) were chosen and carefully prepared so that reactions with the sample are avoided^{38,39}. During preparation, Fe foils and NaCl media and were kept dry, and contact with the atmosphere was minimized to prevent foil oxidation. Carbon from diamond anvils is known to react with Fe at high pressures and temperatures in laser-heated DAC experiments, but generally at much higher temperatures (and longer timescales) than probed in this work^{18,37}. In testing our sample preparation and heating technique in separate *in situ* X-ray diffraction experiments, we ruled out oxidation or reaction with the medium, and confirmed that carbide formation occurs at much higher temperatures and longer heating timescales than we have used here. Thus, our Fe samples should remain very pure at the pressures, temperatures, and timescales of this study. Analysis of the recovered sample from experiments at 58–74 GPa (using electron imaging, energy dispersive scattering for chemical analysis, and a focused ion beam to section the foil at heated regions) found no detectable local enrichment of impurities in the heated areas of the sample, indicating bulk impurity levels well below detection limits ($\leq 0.6 \text{ wt\% C}$, $\leq 0.6 \text{ wt\% O}$, ≤ 100 parts per million Ar), consistent with expectations from X-ray diffraction. Finally, no systematic changes in measured conductivities were observed with heating time, indicating that samples did not undergo any progressive transformation (such as a reaction) that influenced the thermal conductivity.

Model for pressure variation of thermal conductivity. The model used here to estimate pressure variation of thermal conductivity (equation (3)) is based on a formal differentiation of the electronic thermal resistivity ($W_e = 1/k_e$) with respect to density combined with the definition of the Grüneisen parameter ($\gamma = (\partial \ln \theta_D / \partial \ln \rho)_T$, where θ_D is the Debye temperature, and ρ is density), which leads to⁴⁰:

$$\left(\frac{\partial \ln W_e}{\partial \ln \rho} \right)_T = -2\gamma + \left(\frac{\partial \ln C}{\partial \ln \rho} \right)_T \quad (4)$$

where C is a constant containing lattice and band structure information originating from the Bloch–Grüneisen expression. Bohlin⁴¹ finds $(\partial \ln C / \partial \ln \rho)_T$ to be equal to $-1/3$ in ordinary pure metals; the variation of electronic thermal conductivity with pressure can then be expressed in terms of the isothermal bulk modulus (K_T) and the Grüneisen parameter (γ) as equation (3).

The Grüneisen parameter of Fe is fairly well known at high pressure and room temperature: the data of refs 42 and 43 agree well, particularly above 100 GPa. At core conditions (high T), $\gamma(P, T)$ and $K_T(P, T)$ were evaluated using a thermal equation of state of Fe (ref. 27), with $\gamma = \gamma_0 \left(\frac{V}{V_0} \right)^q$, where $\gamma_0 = 1.78$, $q = 0.69$ and $V_0 = 6.73 \text{ cm}^3 \text{ mol}^{-1}$. The P, T description of γ is expressed in a polynomial form:

$$\gamma(P, T) = \frac{a + cP + eT + gP^2 + iT^2 + kPT}{1 + bP + dT + fP^2 + hT^2 + jPT} \quad (5)$$

We described $K_T(P, T)$ by the following equation:

$$K_T(P, T) = K_1 + \frac{K_2 P}{\ln(P)} + \frac{K_3 T}{\ln(T)} \quad (6)$$

All the coefficients for γ and K_T (equations (5) and (6)) are given in Extended Data Table 2.

This model gives good agreement with ϵ Fe electrical resistivity data at lower pressures and ambient temperatures^{5,14}, fits the present thermal conductivity results for ϵ Fe well (Fig. 2), and implies that thermal conductivity is only weakly pressure dependent above 100 GPa, consistent with prior expectations². Thus, our measurements, taken at pressures close to those at the top of Earth's core, should constrain overall core conductivity accurately.

The Lorenz number for ϵ Fe. The temperatures and pressures of our thermal conductivity measurements overlap with those of shock wave electrical resistivity measurements¹⁵, allowing a comparison between the resistivity and thermal conductivity measurements to obtain an empirical value for L .

At 112 GPa, where the most extensive high temperature data set was available in the present results, electrical conductivity was estimated as follows using the data of ref. 15. The two lowest pressure points from that study at 101.1 GPa and 146.7 GPa are solid-state data and so are comparable to the present results; a higher pressure point corresponds to the liquid^{15,18}. First, a temperature for the middle of the three data points (146.7 GPa, 3,357 K) after isentropic release from the initial conditions (173.4 GPa, 3,552 K)—not reported—was estimated from the scaling of release behaviour reported by ref. 15; release temperatures were confirmed by independent calculation using an ϵ Fe equation of state³⁴. The electrical conductivity at 112 GPa is then estimated as $(1.13 \pm 0.11) \times 10^6 \text{ S m}^{-1}$ at 2,332 K, based on a linear interpolation between the solid-state data points, and assuming an uncertainty of $\sim 10\%$, consistent with uncertainties reported for similar measurements¹⁶ and scatter in the data reported by ref. 15. At this temperature in our experiments, $k = 50 \pm 10 \text{ W m}^{-1} \text{ K}^{-1}$ (Fig. 3). The corresponding value of L is then $(1.9 \pm 0.4) \times 10^{-8} \text{ W } \Omega \text{ K}^{-2}$, $22 \pm 16\%$ less than the standard value for a free-electron metal. This correction has a small influence on our analysis. For example, assuming the free-electron value of L , the shock wave results of ref. 15 imply a value of $67 \text{ W m}^{-1} \text{ K}^{-1}$ at 112 GPa and 2,330 K, only slightly above the measured value.

The correction to the standard value of L determined here for ϵ Fe is typical for Fe at various conditions and phases^{7,14,24,44} ($\pm 30\%$) and is similar to other transition metals^{11,45}. In Pt, L is measured¹¹ to deviate from the ideal value by $\pm 30\%$ at temperatures up to 2,000 K. For Mo, deviations of -10% to -30% are predicted at high temperature⁴⁵. The variation of L across transition metals at low temperature alone is large⁴⁶, with values such as in Cu (-9%) and W ($+31\%$).

We note that early shock data on Fe electrical resistivity at high pressures¹⁶ finding systematically higher electrical conductivities compared to later work¹⁵, cannot be considered to agree with our measurements, as an unrealistically large reduction in the Lorenz number would be needed. It has been proposed that spurious values were obtained in the earlier studies¹⁶ at higher pressure ($P > 50$ GPa) owing to insulator–conductor transformation of epoxies used in target construction, an effect avoided in later measurements¹⁵.

Model for temperature variation of thermal conductivity. Equation (2) was selected in consideration of the observed variation of electrical conductivity in ϵ Fe with temperature^{5,14}. Electrical conductivity is modelled as following a relationship:

$$\sigma = \sigma_0 + AT^n \quad (7)$$

where $n = -1$ is typically assumed for metals at high temperatures as in the Bloch–Grüneisen model^{5,7,13,14}. A value closer to $n = -1.3$ has been suggested for Fe at high pressures from resistivity measurements, under both shock and static loading, which probed temperatures and pressures similar to those examined here¹⁴. Similarly, fitting equation (7) to resistivity data under external heating of statically compressed samples⁵, for which temperatures are particularly accurate, yielded values of $n = -1.50 \pm 0.07$, $\sigma_0 = (1.04 \pm 0.46) \times 10^6$, $A = (6.51 \pm 2.2) \times 10^{10}$, for σ in units of siemens per metre and T in units of kelvin (Extended Data Fig. 1a).

Then, considering the Wiedemann–Franz relation (equation (1)), we can write:

$$k = L(T\sigma_0 + AT^{1+n}) \quad (8)$$

leading to the empirical form in equation (2). We chose here $n = -1.5$, though results are not significantly different selecting $n = -1.3$.

Equation (2) is fitted to the present measurements at 112 GPa together with shock wave resistivity data¹⁵, interpolated to 112 GPa as discussed above, and static resistivity data^{5,14} extrapolated to 112 GPa using a double-exponential fit of the form:

$$\frac{1}{\sigma} = \alpha + \beta_1 \exp(\tau_1 P) + \beta_2 \exp(\tau_2 P) \quad (9)$$

An initial fit gave $a = (0.89 \pm 1.33) \times 10^{-3} \text{ W m}^{-1} \text{ K}^{-2}$, $b = 2,040 \pm 140 \text{ W m}^{-1} \text{ K}^{-1/2}$. The linear component of the fit is nearly zero, thus a reasonable simplified version of this model for Fe is:

$$k = b' / \sqrt{T} \quad (10)$$

where $b' = 1,972 \pm 83 \text{ W m}^{-1} \text{ K}^{-1/2}$ (Fig. 3 and Extended Data Fig. 1b).

The model captures a decrease in the thermal conductivity with temperature, which is seen in the present measurements and also implied by the prior resistivity data^{5,14,15} (Fig. 3). In terms of electrical resistivity (Extended Data Fig. 1c), the scaling with temperature obtained by the model compares well with that observed by ref. 5 in ϵ Fe at lower pressures, and shows a similar dependence to that seen in γ Fe (or possibly in the γ – ϵ mixed phase) at high temperatures²¹. It is seen that ϵ Fe up to 112 GPa has higher resistivity than γ Fe (or its mixed phase) at lower pressure (Extended Data Fig. 1c), consistent with our experimental observation of higher thermal conductivity in γ Fe compared to ϵ Fe in the low-pressure region (Fig. 2).

We note that the minimum measured thermal conductivity is in close agreement with values expected at traditional resistivity saturation⁵ (Extended Data Fig. 1b); however, as resistivity saturation in Fe at extremes has not been clearly confirmed by theoretical studies and since available saturation models⁵ cannot satisfactorily describe the data, we conclude that at present there is no reason to adopt that resistivity saturation has occurred. Assuming it has, then ϵ Fe at temperatures above $\sim 3,000$ K is saturation-dominated, such that thermal conductivities at core conditions would be somewhat higher (60 – $80 \text{ W m}^{-1} \text{ K}^{-1}$) than assessed by the present modelling; however, this upper bound on conductivity is still low compared to many prior estimates, and would not substantially alter our main conclusions.

Error assessment in the thermal conductivity determination. The laser-heated DAC in combination with numerical simulations has been shown to be a promising tool for studying heat transfer at high pressures and temperatures^{11,12,26,30,47–50}. This approach requires a detailed understanding of heat transfer in the DAC, including quantitative relationships between the temperature distribution, pressure chamber geometries and sample physical properties.

Finite-element model fits to temperature histories were generally performed using a manual adjustment of model parameters. This approach was evaluated against a Levenberg–Marquardt least-squares minimization of the finite-element model variables (Extended Data Fig. 5). This automatic optimization was able to improve fit quality but the improvement was not statistically significant. Furthermore, as a good initial guess was required, this additional step only added to the processing time, and was therefore not used for all data sets.

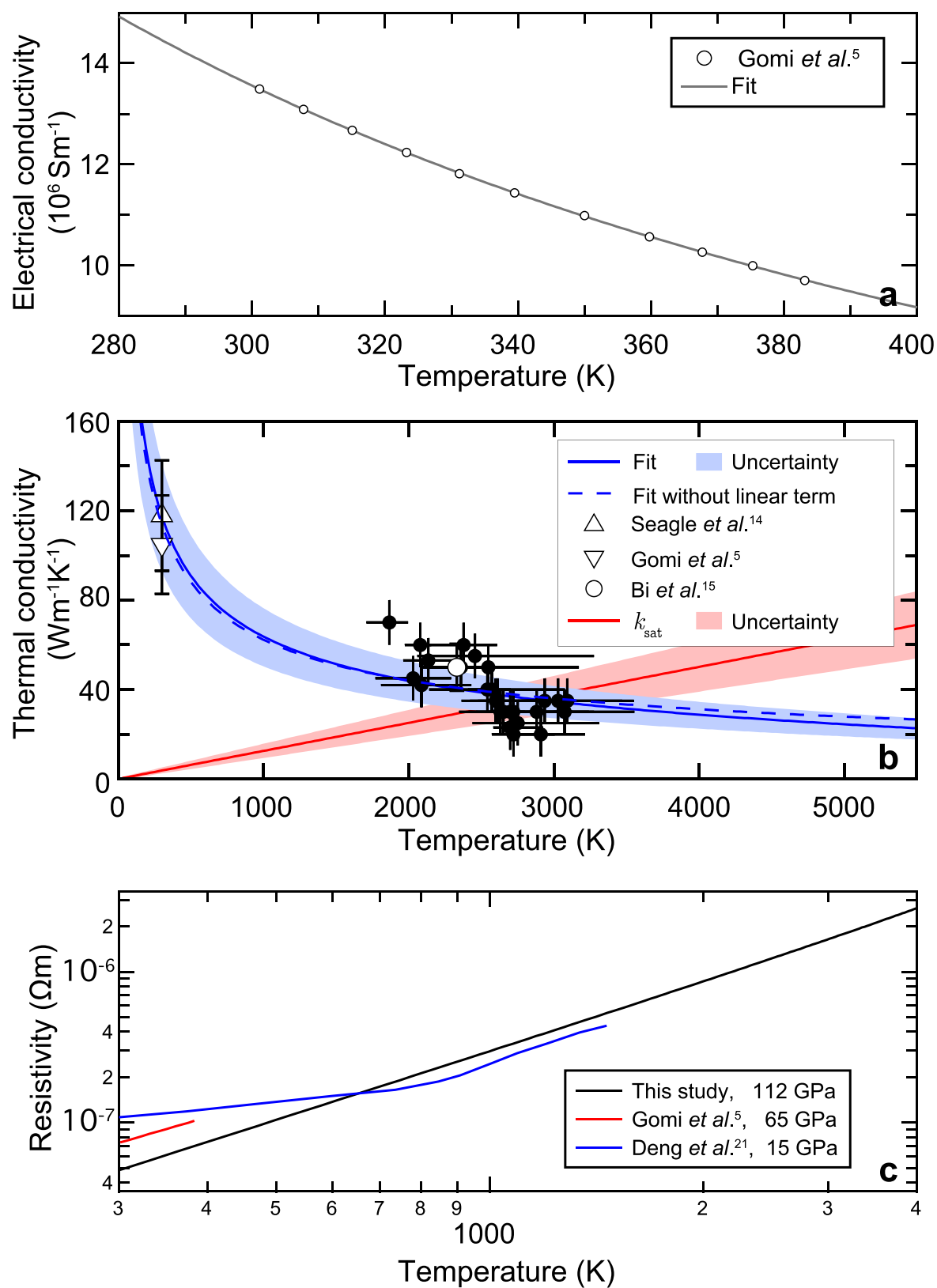
In the present study, all input parameters in modelling were carefully examined for their effect on the determination of sample thermal conductivity (Extended Data Figs 2 and 3). Uncertainties in the input parameters (such as pressure chamber geometry) were in this way included in our overall uncertainty determination for k . The heat capacity C_p of the pressure medium has a negligible effect (Extended Data Fig. 3a). For C_p of Fe we derived a range of values of 500 – $700 \text{ J kg}^{-1} \text{ K}^{-1}$ from equations of state for ϵ Fe (refs 34 and 51) and other estimates⁵². Within this range, the resulting sample k is unaffected (Extended Data Fig. 3b). Thermal conductivity of the diamond anvils, temperature dependence of thermal conductivity of the pressure medium, and smaller or larger laser beam size (by about 13%) also have negligible effects on the sample k (Extended Data Fig. 3c–e). Sample and insulation layer thicknesses, on the other hand, contribute to the uncertainty in sample k : an approximately $\pm 20\%$ change in thicknesses leads to $\pm 7 \text{ W m}^{-1} \text{ K}^{-1}$ changes in sample k (Extended Data Fig. 3f–i). We assume a constant value of k for the foil in our simulations, but no significant change in results is obtained using a temperature-dependent k (Extended Data Fig. 3j).

To check potential couplings between the uncertainties in the input parameters, we have also propagated uncertainty in our input parameters in a more rigorous manner using a Monte Carlo approach (Extended Data Fig. 6). To do this, we considered only parameters which were identified as having a first-order impact on the measurements: the thicknesses of the medium on both sides of the sample, and the sample thickness. We performed 64 Monte Carlo samples within the Gaussian probability distributions of the thickness parameters, given standard deviations of 30% in each, for a representative experiment at 130 GPa (see Extended Data Fig. 6a). For each sampling, the data was fitted automatically (Extended Data Fig. 5) to determine the two thermal conductivities and the powers for the three lasers (Extended Data Fig. 6b). The distribution in the values for k of Fe has a standard deviation comparable to our single-point error (Extended Data Fig. 6d).

While suitably sensitive to the thermal conductivity of the foil, our measurements are less sensitive to the thermal conductivity of the insulating medium, which is included as a variable in fitting (usually as a constant) but which had values more sensitive to the assumed sample geometry (thickness of the insulation layers), laser beam diameter and laser power. Thus, the conductivities of

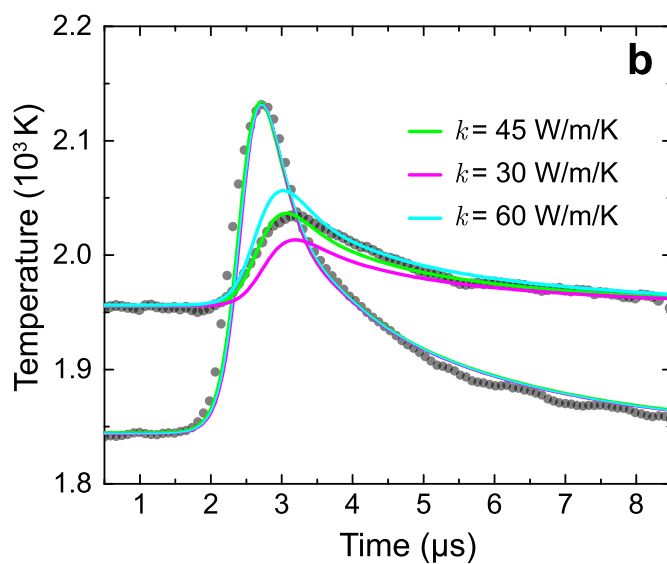
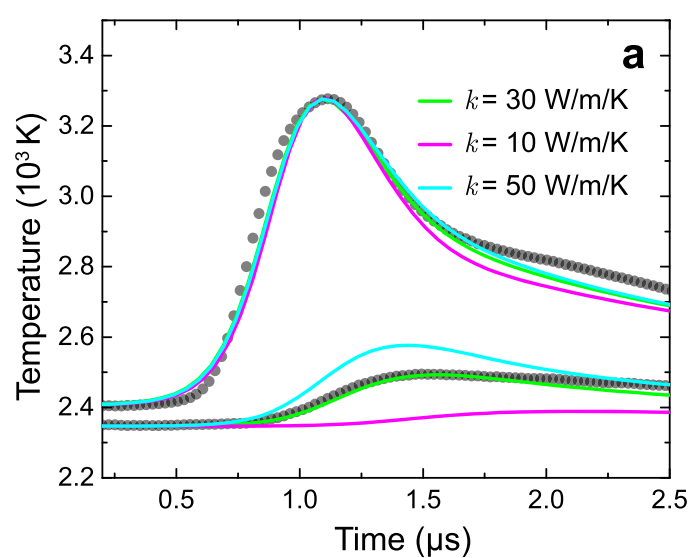
insulating media are not reported, as they are not robustly determined by our approach. For Ar, the values of k obtained in the fits were generally in the range of 50–100 W m⁻¹ K⁻¹, consistent with previously reported values⁴⁹.

30. Montoya, J. A. & Goncharov, A. F. Finite element calculations of the time dependent thermal fluxes in the laser-heated diamond anvil cell. *J. Appl. Phys.* **111**, 112617 (2012).
31. Johannsen, P. G. Refractive index of the alkali halides. 1. Constant joint density of states model. *Phys. Rev. B* **55**, 6856–6864 (1997).
32. Grimsditch, M., Letoullec, R., Polian, A. & Gauthier, M. Refractive index determination in diamond anvil cells: results for argon. *J. Appl. Phys.* **60**, 3479–3481 (1986).
33. Chen, B. *et al.* Elasticity, strength, and refractive index of argon at high pressures. *Phys. Rev. B* **81**, 144110 (2010).
34. Dewaele, A. *et al.* Quasihydrostatic equation of state of iron above 2 Mbar. *Phys. Rev. Lett.* **97**, 215504 (2006).
35. Sommerfeld, A. Zur Elektronentheorie der Metalle auf Grund der Fermischen Statistik. *Z. Phys.* **47**, 1–32 (1928).
36. Parker, W. J., Jenkins, R. J., Abbott, G. L. & Butler, C. P. Flash method of determining thermal diffusivity, heat capacity, and thermal conductivity. *J. Appl. Phys.* **32**, 1679 (1961).
37. Tateno, S., Hirose, K., Ohishi, Y. & Tatsumi, Y. The structure of iron in Earth's inner core. *Science* **330**, 359–361 (2010).
38. Shen, G., Prakapenka, V. B., Rivers, M. L. & Sutton, S. R. Structure of liquid iron at pressures up to 58 GPa. *Phys. Rev. Lett.* **92**, 185701 (2004).
39. Goncharov, A. F. *et al.* X-ray diffraction in the pulsed laser heated diamond anvil cell. *Rev. Sci. Instrum.* **81**, 113902 (2010).
40. Ross, R. G., Andersson, P., Sundqvist, B. & Backstrom, G. Thermal conductivity of solids and liquids under pressure. *Rep. Prog. Phys.* **47**, 1347 (1984).
41. Bohlin, L. Thermal conduction of metals at high pressure. *Solid State Commun.* **19**, 389–390 (1976).
42. Sharma, S. K. Debye temperature of hcp iron at extreme compression. *Solid State Commun.* **149**, 2207–2209 (2009).
43. Dubrovinsky, L. S., Saxena, S. K., Dubrovinskaia, N. A., Rekhi, S. & Le Bihan, T. Gruneisen parameter of ϵ -iron up to 300 GPa from in-situ X-ray study. *Am. Mineral.* **85**, 386–389 (2000).
44. Van Zytveld, J. Electrical resistivities of liquid transition metals. *J. Phys. Coll.* **41**, C8-503-C8-506 (1980).
45. French, M. & Mattsson, T. R. Thermoelectric transport properties of molybdenum from ab-initio simulations. *Phys. Rev. B* **90**, 165113 (2014).
46. Kittel, C. *Introduction to Solid State Physics* 8th edn (John Wiley & Sons, 2005).
47. Panero, W. R. & Jeanloz, R. Temperature gradients in the laser-heated diamond anvil cell. *J. Geophys. Res. Solid Earth* **106**, 6493–6498 (2001).
48. Kiefer, B. & Duffy, T. S. Finite element simulations of the laser-heated diamond-anvil cell. *J. Appl. Phys.* **97**, 114902 (2005).
49. Goncharov, A. F. *et al.* Thermal conductivity of argon at high pressures and high temperatures. *J. Appl. Phys.* **111**, 112609 (2012).
50. Beck, P. *et al.* Measurement of thermal diffusivity at high pressure using a transient heating technique. *Appl. Phys. Lett.* **91**, 181914 (2007).
51. Yamazaki, D. *et al.* P-V-T equation of state for ϵ -iron up to 80 GPa and 1900 K using the Kawai-type high pressure apparatus equipped with sintered diamond anvils. *Geophys. Res. Lett.* **39**, L20308 (2012).
52. Hirose, K., Labrosse, S. & Hernlund, J. Composition and state of the core. *Annu. Rev. Earth Planet. Sci.* **41**, 657–691 (2013).



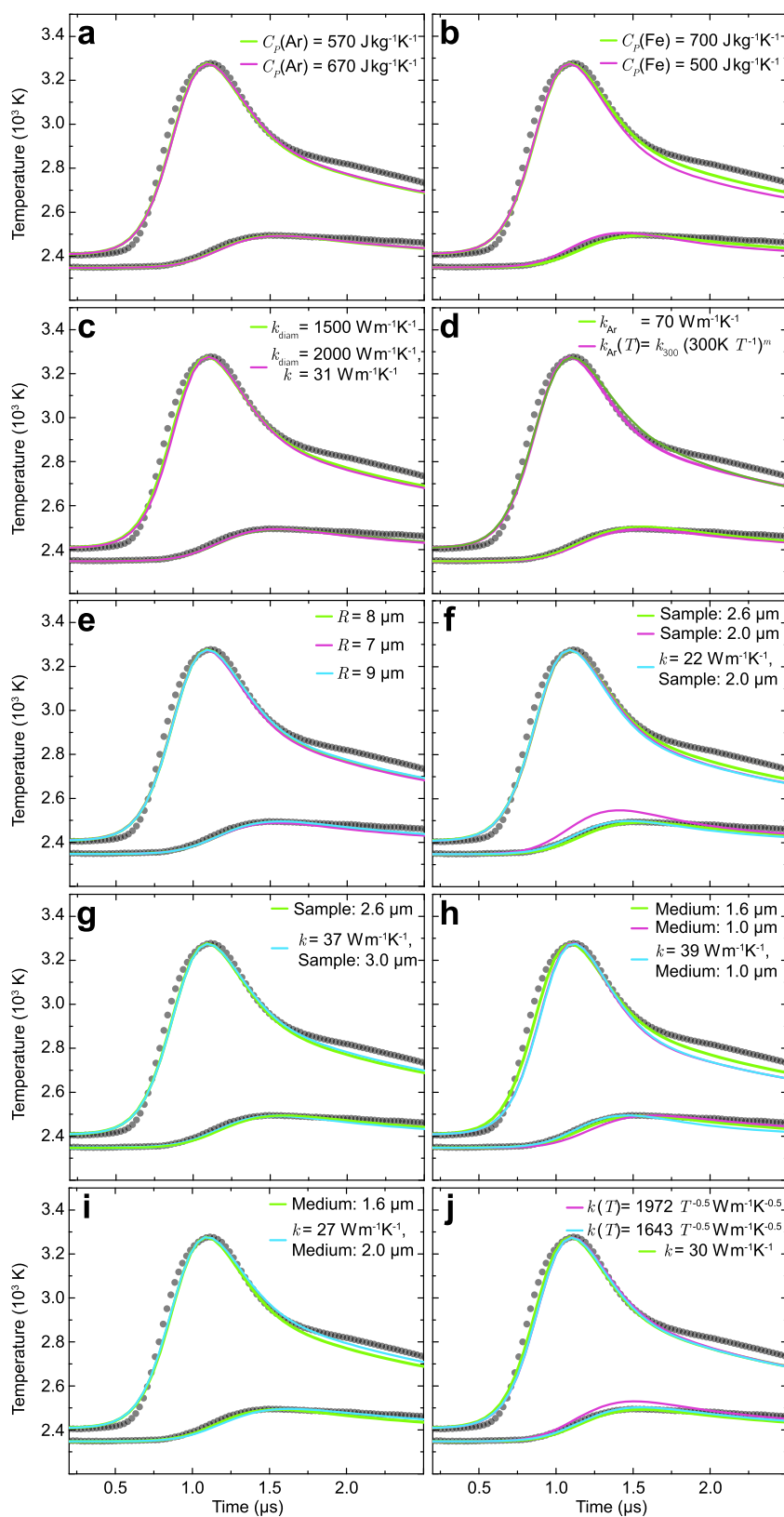
Extended Data Figure 1 | High-temperature transport properties of Fe. **a**, Graph of the electrical conductivity⁵ as a function of temperature of ϵ Fe at 65 GPa and model fit (to equation (7)). **b**, Thermal conductivity temperature dependence at 112 GPa. The model fit (to equation (2), solid line) and a 20% uncertainty envelope are in blue; the model fit without linear term (to equation (10)) is a dashed blue line. Present data are solid

circles and data derived from prior electrical resistivity measurements^{5,14,15} are open symbols (see Fig. 3). The red band is the minimum thermal conductivity assuming resistivity saturation⁵. **c**, Electrical resistivity at several pressures, for multiple phases at 15 GPa (blue)²¹, and the ϵ phase at 65 GPa (red)⁵ and 112 GPa (this study, black).



Extended Data Figure 2 | Comparison between measurements and models for different values of thermal conductivity. Data for pulsed and opposite sides of the foil are dots; the larger temperature excursion is on the pulsed side. Green, magenta and cyan curves are simulations with

different values of sample k , all other parameters being held constant. The data sets at 112 GPa (a) and 130 GPa (b) have been measured using 3- μ s and 10- μ s sweep windows, respectively.

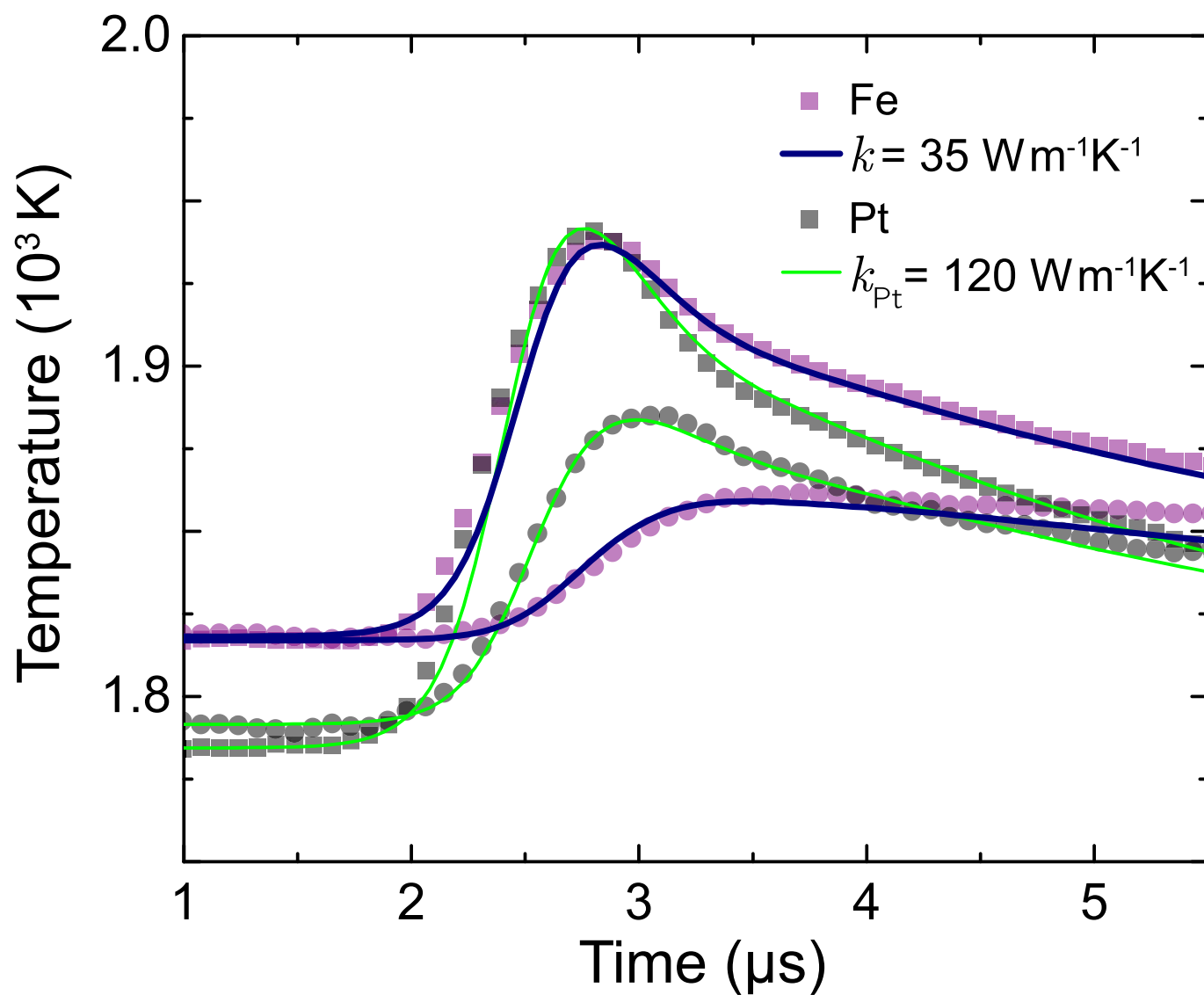


Extended Data Figure 3 | See next page for caption.

Extended Data Figure 3 | Tests of the sensitivity of finite-element model results to input parameters for an example run at 112 GPa.

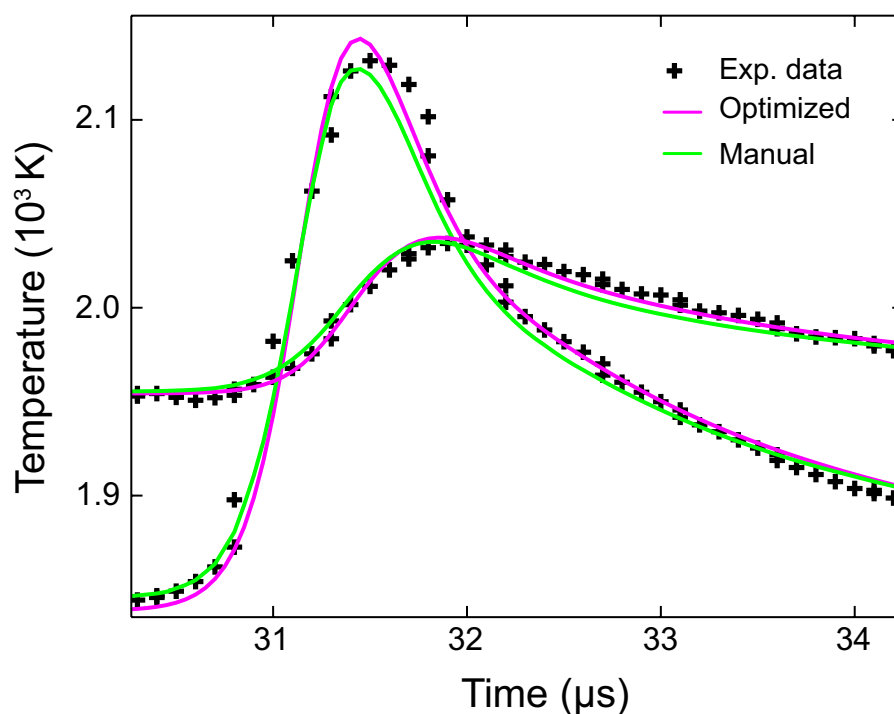
This experiment shows a large amplitude of temperature modulation that accentuates the effects of parameter changes. A best-fit value of $k = 30 \text{ W m}^{-1} \text{ K}^{-1}$, obtained using parameters listed in Extended Data Table 1, is obtained from these model fits unless stated otherwise. **a**, Effect of heat capacity of the Ar pressure medium. Uncertainty in medium C_p has no effect on k of the sample. **b**, Effect of heat capacity of the sample. Temperature profiles for two values of C_p of Fe ($500 \text{ J kg}^{-1} \text{ K}^{-1}$ and $700 \text{ J kg}^{-1} \text{ K}^{-1}$) indicate that results are only weakly affected by the uncertainty in C_p for Fe. **c**, Change in the thermal conductivity value of diamond anvils from $1,500 \text{ W m}^{-1} \text{ K}^{-1}$ to $2,000 \text{ W m}^{-1} \text{ K}^{-1}$ requires an increase in thermal conductivity of the sample from $30 \text{ W m}^{-1} \text{ K}^{-1}$ to $31 \text{ W m}^{-1} \text{ K}^{-1}$. **d**, Effect of using a T -dependent k of the medium. After ref. 49, a dependence $k(T) = k_{300}(300/T)^m$ is used, where k_{300} is the 300-K conductivity, T is in kelvin, and m is an exponent (of order 1); k_{300} ($300 \text{ W m}^{-1} \text{ K}^{-1}$) is extrapolated from prior results at lower pressure⁴⁹

and m (0.7) is fitted to the present data. No change in sample k is indicated using this or any other $k(T)$ model we tested for the media. **e**, Laser beam radius change of $\pm 13\%$ does not affect the temperature noticeably. **f**, A sample thinner by 23% (reduced from $2.6 \mu\text{m}$ to $2.0 \mu\text{m}$) would require a lower sample k of $22 \text{ W m}^{-1} \text{ K}^{-1}$. **g**, A sample thicker by 15% (increased from $2.6 \mu\text{m}$ to $3.0 \mu\text{m}$) would require an increased sample k of $37 \text{ W m}^{-1} \text{ K}^{-1}$. **h**, The insulation layer was decreased on both sides by 38%, from $1.6 \mu\text{m}$ to $1.0 \mu\text{m}$. Sample k had to increase to $39 \text{ W m}^{-1} \text{ K}^{-1}$. **i**, The insulation layer was increased on both sides by 25%, from $1.6 \mu\text{m}$ to $2.0 \mu\text{m}$. Sample k had to decrease to $27 \text{ W m}^{-1} \text{ K}^{-1}$. **j**, Effect of including T dependence of sample k in models. The temperature profile calculated using our global fit at 112 GPa (equation (2)) is shown as a magenta line; this dependence scaled within its uncertainty (reduced by a factor of 0.83) to improve the fit is shown as a cyan line. The resulting sample k varies between $24 \text{ W m}^{-1} \text{ K}^{-1}$ and $35 \text{ W m}^{-1} \text{ K}^{-1}$ in the T range of the experiment; the estimate assuming constant sample k is the average of these values.



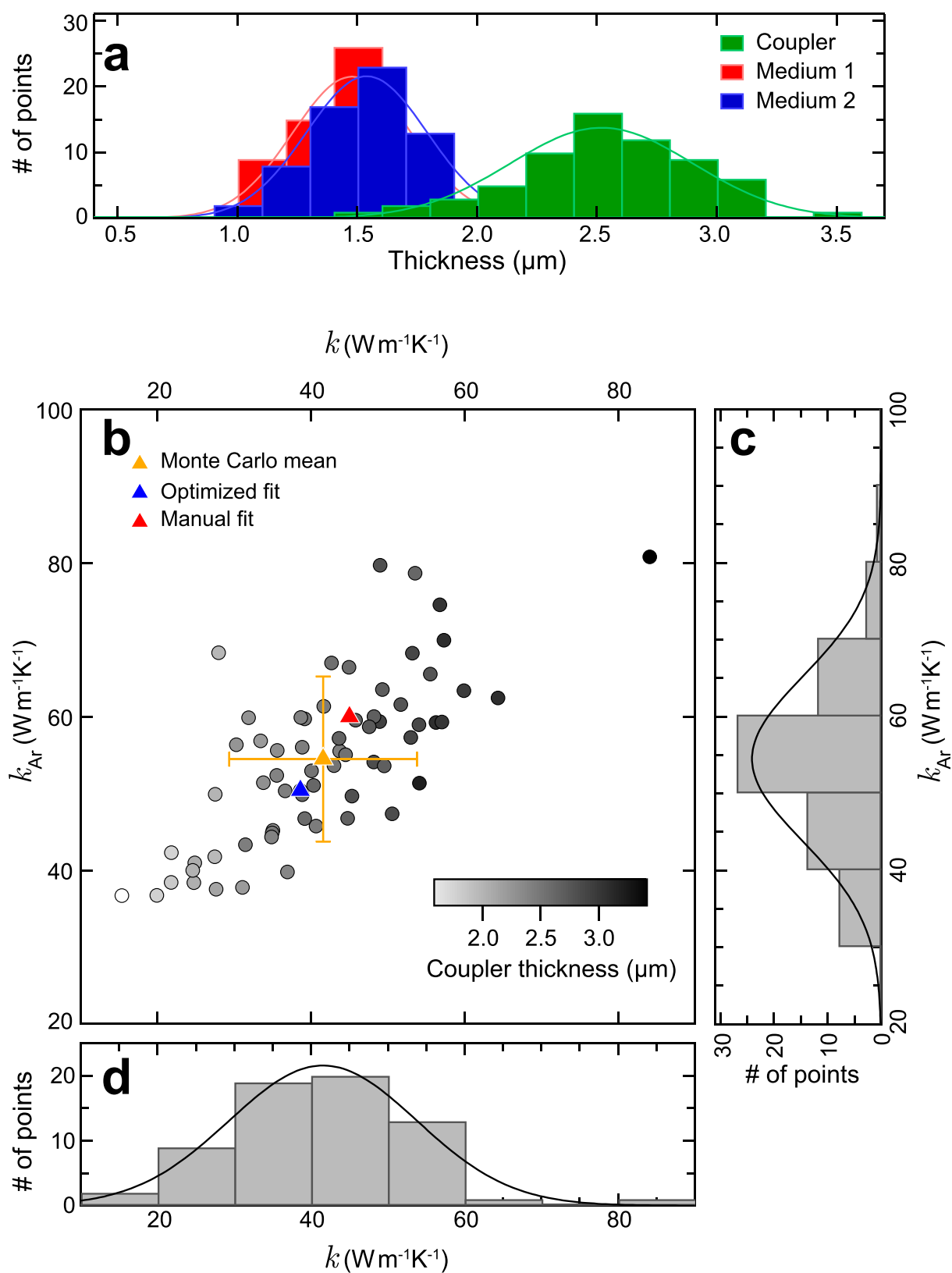
Extended Data Figure 4 | Comparison of data on Fe and Pt at 48 GPa for an identical sample configuration. The data clearly show slower propagation of heat across the Fe foil compared to Pt (ref. 11), as given by the half-rise time τ . This observation directly shows that thermal

diffusivity $\kappa = (k/\rho C_p)$ of Fe is much less than Pt, since^{11,36} $\kappa \propto 1/\tau$. Similarly, the smaller amplitude of the perturbation upon opposite surface arrival indicates a smaller k in Fe than in Pt.



Extended Data Figure 5 | Comparison between manual and automatic optimization results for an experiment at 130 GPa. The manual approach, used as our primary fitting method, was based on an adjustment of model parameters by hand within a precision of $\sim 5 \text{ W m}^{-1} \text{ K}^{-1}$, giving $k = 45 \text{ W m}^{-1} \text{ K}^{-1}$ and $k_{\text{Ar}} = 60 \text{ W m}^{-1} \text{ K}^{-1}$ as the best fit. The automatic result is the best fit based on a Levenberg–Marquardt least-squares

minimization of model parameters, yielding $k = 38.6 \text{ W m}^{-1} \text{ K}^{-1}$ and $k_{\text{Ar}} = 50.4 \text{ W m}^{-1} \text{ K}^{-1}$. The automatic optimization obtained a better least-squares fit (χ^2 improved by 23%); however, the difference in k is not statistically significant.



Extended Data Figure 6 | Monte Carlo analysis of error coupling in thickness uncertainties and effect on thermal conductivities, for the 130-GPa data set shown in Extended Data Fig. 5. a, Histogram showing randomly sampled thicknesses (upper and lower medium, and foil) in Gaussian probability distributions with standard deviation 30%. **b,** Thermal conductivities for Ar and Fe for 64 samples. The greyscale

refers to the value of the coupler thickness, showing the correlation between high values for k and thicker coupler. The results of fits shown in Extended Data Fig. 5 are blue and red triangles, while the mean and one standard deviation found from the spread of sampled thermal conductivities is the orange triangle. **c** and **d** are histograms showing the distribution of thermal conductivities in **b**.

Extended Data Table 1 | Input parameters used for the finite-element modeling

P (GPa)	medium	sample thickness (μm)	pulsed side thickness (μm)	opposite side thickness (μm)	medium density (kg m^{-3})	C_p medium ($\text{J kg}^{-1}\text{K}^{-1}$)	iron density (kg m^{-3})	C_p iron ($\text{J kg}^{-1}\text{K}^{-1}$)
35	NaCl	3.0	8.0	7.0	3630	$\begin{cases} 748 + 0.34 T, & T < 1000\text{K} \\ 1103, & T > 1000\text{K} \end{cases}$	9602	450
48	NaCl	2.9	7.4	6.7	3911	$\begin{cases} 748 + 0.34 T, & T < 1000\text{K} \\ 1103, & T > 1000\text{K} \end{cases}$	9929	450
58	Ar	2.9	1.5	6.5	4539	570	10174	700
74	Ar	2.8	1.0	6.4	4800	570	10476	700
88	Ar	2.7	1.7	1.7	5057	570	10800	700
112	Ar	2.6	1.6	1.6	5326	570	11225	700
130	Ar	2.5	1.5	1.5	5550	570	11590	700

Extended Data Table 2 | Coefficients for the Grüneisen parameter and the isothermal bulk modulus used to estimate pressure variation of thermal conductivity

Coefficient for Grüneisen parameter		Coefficient for K_T	
<i>a</i>	1.76×10^0	K_1	97.50
<i>b</i>	2.04×10^{-2}	K_2	25.77
<i>c</i>	2.90×10^{-2}	K_3	-0.26
<i>d</i>	-1.32×10^{-4}		
<i>e</i>	-1.87×10^{-4}		
<i>f</i>	3.90×10^{-5}		
<i>g</i>	3.42×10^{-5}		
<i>h</i>	2.55×10^{-9}		
<i>i</i>	3.05×10^{-9}		
<i>j</i>	-5.10×10^{-7}		
<i>k</i>	-4.37×10^{-7}		

The industrial melanism mutation in British peppered moths is a transposable element

Arjen E. van't Hof^{1*}, Pascal Campagne^{1*}, Daniel J. Rigden¹, Carl J. Yung¹, Jessica Lingley¹, Michael A. Quail², Neil Hall¹, Alistair C. Darby¹ & Ilik J. Saccheri¹

Discovering the mutational events that fuel adaptation to environmental change remains an important challenge for evolutionary biology. The classroom example of a visible evolutionary response is industrial melanism in the peppered moth (*Biston betularia*): the replacement, during the Industrial Revolution, of the common pale *typica* form by a previously unknown black (*carbonaria*) form, driven by the interaction between bird predation and coal pollution¹. The *carbonaria* locus has been coarsely localized to a 200-kilobase region, but the specific identity and nature of the sequence difference controlling the *carbonaria*–*typica* polymorphism, and the gene it influences, are unknown². Here we show that the mutation event giving rise to industrial melanism in Britain was the insertion of a large, tandemly repeated, transposable element into the first intron of the gene *cortex*. Statistical inference based on the distribution of recombinant *carbonaria* haplotypes indicates that this transposition event occurred around 1819, consistent with the historical record. We have begun to dissect the mode of action of the *carbonaria* transposable element by showing that it increases the abundance of a *cortex* transcript, the protein product of which plays an important role in cell-cycle regulation, during early wing disc development. Our findings fill a substantial knowledge gap in the iconic example of microevolutionary change, adding a further layer of insight into the mechanism of adaptation in response to natural selection. The discovery that the mutation itself is a transposable element will stimulate further debate about the importance of ‘jumping genes’ as a source of major phenotypic novelty³.

Ecological genetics, the study of polymorphism and fitness in natural populations, has been revitalised through the application of next-generation sequencing technology to open up what were previously treated as genetic black boxes^{4,5}. Growing appreciation of the loci and developmental networks that generate adaptive phenotypic variation⁶ promises to answer fundamental questions about the genetic architecture of adaptation, such as the prevalence of genomic hotspots for adaptation⁷, the relative contributions of major- and minor-effect mutations⁸, and the structural nature and mode of action of beneficial mutations⁹. Characterizing the identity and origin of functional sequence polymorphisms provides the explicit link between the mutation process and natural selection. In this context, while industrial melanism in the peppered moth has retained its appeal as a graphic example of the spread of a novel mutant rendered favourable by a major change in the environment, the crucial piece of the puzzle that has been missing is the molecular identity of the causal mutation(s)¹⁰.

A combined linkage and association mapping approach previously localized the *carbonaria* locus to a <400-kb region orthologous to *Bombyx mori* chromosome 17 (loci *b*–*d*)². Thirteen genes and two microRNAs occur within this interval, none of which was known to be

involved in wing pattern development or melanization. By extending the association mapping approach to a larger population sample and more closely spaced genetic markers (see Methods), we narrowed the *carbonaria* candidate region to about 100 kb (Fig. 1a). The candidate region resides entirely within the span of one gene — the orthologue of *Drosophila cortex* (*cort*), the only known function of which is as a cell-cycle regulator during meiosis¹¹. In *B. betularia*, *cortex* consists of eight non-first exons, multiple alternative first exons (of which only two, 1A and 1B, are strongly expressed in developing wing discs), and a very large first intron (Fig. 1b).

The rapid spread of *carbonaria* gave rise to strong linkage disequilibrium², such that many sequence variants are associated with the *carbonaria* phenotype. This poses a challenge for isolating the specific causal variant(s). We reasoned that if the *carbonaria* mutation arose on an ancestral *typica* haplotype², the hitchhiking variants should in principle also be present at some frequency within the *typica* population, leaving the causal variants as the only ones unique to *carbonaria*. High-quality contiguous reference sequences were assembled from tiled bacterial artificial chromosome (BAC) and fosmid clones, resulting in one *carbonaria* and three different *typica* core haplotypes (see Methods and Extended Data Fig. 1). Alignment of these sequences (Supplementary Data 1) revealed 87 melanization candidate polymorphisms (Fig. 1b and Supplementary Table 1), concentrated within the large first intron of *cortex* (69–91 kb, depending on haplotype). Eighty-five candidates were eliminated using an increasing number of *typica* individuals to exclude rare variants. A single nucleotide polymorphism (*carbonaria_candidate_25*) was eventually excluded on the basis of one individual out of 283 *typica*, leaving a very large insert (*carbonaria_candidate_45*) as the only remaining candidate.

The insert was found to be present in 105 out of 110 fully black moths (wild caught in the UK since 2002) and absent in all (283) *typica* tested (see Methods and Extended Data Fig. 2). Consistent with local *carbonaria* morph frequencies of 10–30% (ref. 12), 2 out of 105 individuals were homozygous for the *carbonaria* insert. Five individuals that were morphologically indistinguishable from *carbonaria* did not possess the *carbonaria* insert; they do not present any strong haplotype association based on this set of candidate loci but do all differ from the core *carbonaria* haplotype at many positions. Our interpretation is that these individuals are hetero- or homozygous for the most extreme of the *insularia* alleles (intermediate phenotypes), which are known to occasionally produce *carbonaria*-like phenotypes^{13,14} and segregate as alleles of the *carbonaria* locus in classical genetics crosses¹⁴. Conversely, none of the genotyped *insularia* morphs (31 individuals, covering the full spectrum of variation from *i*₁ to *i*₃ (ref. 14)) contains the *carbonaria* insert (Extended Data Fig. 2). We conclude that the large insert is the *carbonaria* mutation.

The *carbonaria* insert is 21,925 nucleotides long and is composed of a roughly 9-kb essentially non-repetitive sequence (except for

¹Institute of Integrative Biology, University of Liverpool, Biosciences Building, Crown Street, Liverpool L69 7ZB, UK. ²Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridgeshire CB10 1SA, UK.

*These authors contributed equally to this work.

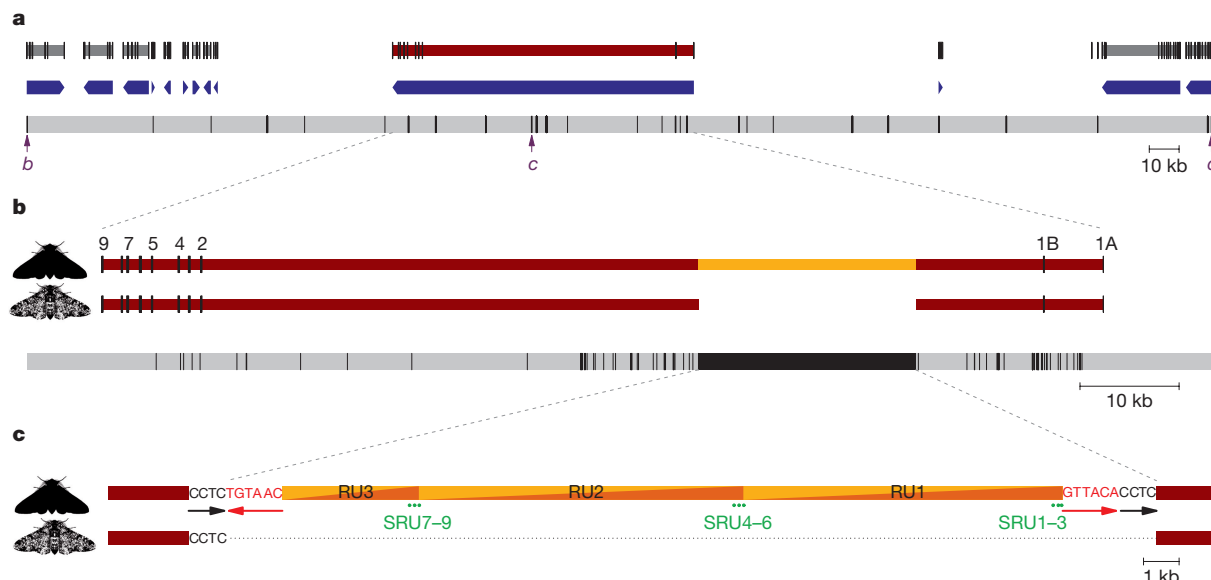


Figure 1 | The *carbonaria* candidate region, and the position and structure of the *carbonaria* mutation. **a**, Approximately 400-kb candidate region (bounded by marker loci *b* and *d* (ref. 2)) indicating gene content and genotyping positions (vertical lines in the continuous grey bar). Intron–exon structure and orientation are illustrated separately for each gene (annotated in GenBank accession KT182637). **b**, Refined candidate region including candidate polymorphisms (lines on the grey bar). The intron–exon structure of *cortex* is shown for *carbonaria* (black moth) and *typica* (speckled moth), highlighting the presence of a large (22 kb) indel (orange) within the first intron. Exons 1A and 1B are alternative transcription starts followed by the shared exons 2–9. **c**, The only exclusive

carbonaria–typica polymorphism within the candidate region. The structure of the insert, shown in the *carbonaria* sequence, corresponds to a class II DNA transposon, with direct repeats resulting from target site duplication (black nucleotides) next to inverted repeats (red nucleotides). *Typica* haplotypes (lower sequence) lack the 4-base target site duplication, the inverted repeats and the core insert sequence. The transposon consists of ~9 kb tandemly repeated two and one-third times (repeat unit (RU)1–RU3), with three short tandem subrepeat units (green dots, SRU1–SRU9) within each repeat unit. Moth images were created from photographs taken by A.E.v.H.

approximately 370 nucleotides at the repeat unit junctions) that is tandemly repeated approximately two and one-third times, with only minor differences among the repeats (Fig. 1c). The insert bears the hallmark of a class II (DNA cut-and-paste) transposable element: short inverted repeats (6 bp) and duplication of the (4-bp) target site present in *typica* haplotypes (Extended Data Fig. 3). We estimate that there are approximately 255 and 60 genomic copies, respectively, of the 9-kb *carbonaria* transposable element (*carb-TE*) repeat unit and repeat unit junctions, implying that there are relatively few genomic copies of the complete *carb-TE*. No nucleotide or translated BLAST hits were found in any relevant database, with the exception of *B. betularia* RNA-sequencing (RNA-seq) reads (NCBI: SRX371328), indicating that the *carb-TE* repeat unit is *Biston*-specific.

To examine patterns of recombination, which provide insight into the evolutionary dynamics of a chromosomal region, we genotyped the same 105 *carbonaria* and a sub-set of 37 *typica*, plus 35 *insularia*, at 119 polymorphic loci within 28 PCR fragments distributed across 200 kb either side of the *carb-TE* (Fig. 1a). Diploid genotypes were phased, and the resulting haplotypes divided into those with and without the *carb-TE*. The sequence identity of the ancestral *carbonaria* haplotype, whose core was known from the BAC and fosmid work, was extended by assigning allelic state at each marker locus to ancestral *carbonaria* or *typica/insularia*. Fifty per cent of *carb-TE* haplotypes had retained the ancestral *carbonaria* haplotype across the full 400-kb window, and the remainder showed varying degrees of recombination with *typica* haplotypes on one or both sides of the causal mutation (Fig. 2a). The recent selective sweep¹⁵ is reflected by declining linkage disequilibrium between the *carbonaria* locus and marker loci with increasing genetic distance (Fig. 2b). The tenure of the *carb-TE* has been transient, having declined from ~99% to less than 5% in its industrial heartland since 1970 (ref. 16). It has nevertheless left a substantial trace of its former abundance in the form of ancestral *carbonaria* haplotype blocks introgressed into *typica* and *insularia* haplotypes, consistent with the simulation-based expectation (Fig. 2c).

The first reported sighting of the *carbonaria* form is generally regarded as having occurred in 1848 in Manchester¹, although the wording of the record implies that it was rare but not completely unknown at this time. Establishing how long before this date the *carbonaria* mutation occurred is complicated because it could have existed undetected at a low frequency for hundreds of years (Supplementary Methods). Our approach to this problem was to infer the age of the mutation event independently by considering the erosion of the ancestral *carbonaria* haplotype due to genetic recombination and mutation. One million simulated time trajectories of the *carbonaria* phenotype were randomly drawn according to their fit to historical frequency data (Extended Data Fig. 4). Based on these trajectories, recombination patterns were simulated using an empirical estimate of recombination rate and compared to the observed recombination pattern of the *carbonaria* haplotypes. The probability density for the date of the *carb-TE* mutation event (Fig. 2d) is highly skewed (median, 1763; interquartile range, 1681–1806) with a maximum likelihood at 1819, a date highly consistent with a detectable frequency being achieved in the mid-1840s.

The position of the *carb-TE* suggests that its effect on melanization is achieved by altering the expression of *cortex* through one of several potential mechanisms¹⁷ (incorporation of any part of the *carb-TE* into *cortex* transcripts has been excluded). *Biston cortex* is characterized by numerous splice isoforms and alternative first exons; we focus on the population of transcripts initiated by exons 1A and 1B, as the other first exons are absent or only weakly expressed in *Biston* wing discs, and did not exhibit morph-specific differences (Extended Data Fig. 5). The global pattern of splice isoforms showed neither consistent presence or absence nor crude relative abundance differences among morphs for any developmental stage (Extended Data Figs 6, 7). Cumulative expression across all isoforms (Fig. 3a) increases by an order of magnitude between the sixth larval instar (La6) and day 4 prepupa (Cr4), coinciding with a phase of rapid wing disc morphogenesis (Fig. 3b), and falls back to a low level by day 6 prepupa (Cr6) with

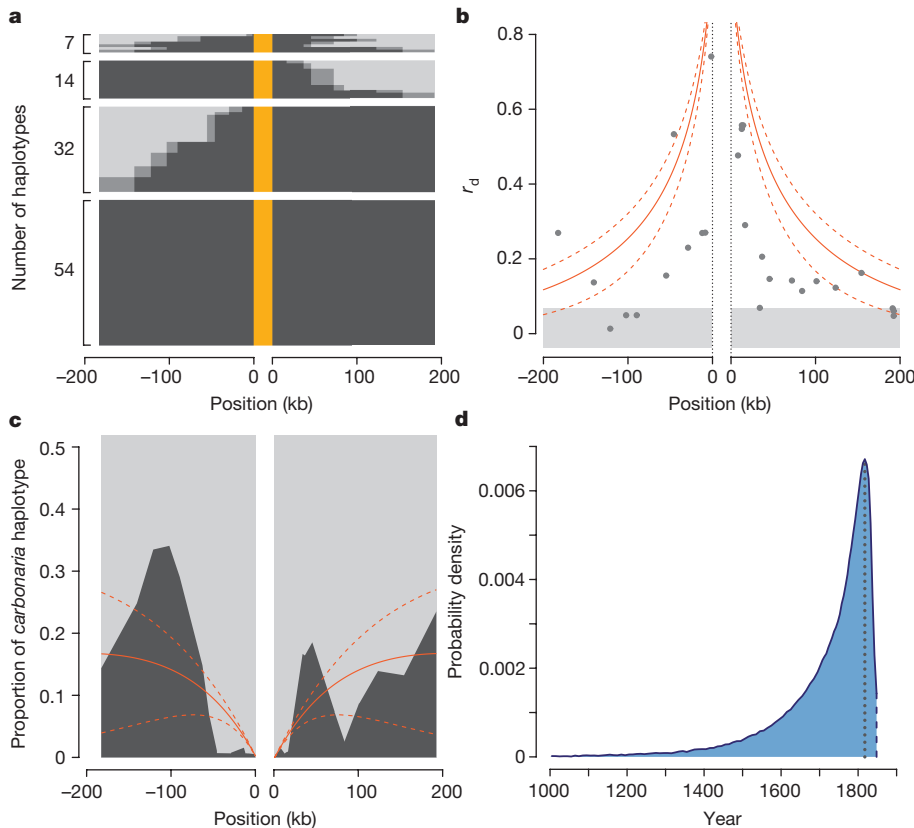


Figure 2 | Recombination pattern and ageing of the *carb*-TE mutation. **a**, Nearest recombination between *carbonaria* (*carb*-TE present (orange)) and non-*carbonaria* (*typica* and *insularia* (light grey)) haplotypes ($n = 107$), 200 kb either side of the *carb*-TE (at position 0). Dark grey areas indicate boundaries within which recombination occurred. **b**, Multilocus linkage disequilibrium (r_d) across the same sequence window among *carbonaria* and non-*carbonaria* haplotypes. Grey area indicates the widest 99% confidence region, across loci, for the null hypothesis ($r_d \approx 0$). Red lines represent the simulation-based upper bound under the extreme assumption that all alleles defining the *carbonaria* haplotype were initially exclusive to it (mean and 90% interval). **c**, Introgression of the ancestral *carbonaria* haplotype (black) into non-*carbonaria* haplotypes (grey; *carb*-TE absent ($n = 144$)). Red lines represent the simulation-based expectations (mean and 90% interval). **d**, Probability density for the age of the *carb*-TE mutation inferred from the recombination pattern in the *carbonaria* haplotypes (maximum density at 1819 shown by dotted line; first record of *carbonaria* in 1848 shown by dashed line).

no clear difference among morphs (t/t versus c/t , $P > 0.5$). To exclude interference by potentially non-functional isoforms, we targeted full transcripts only, starting with either 1A or 1B. The abundance of the 1B full transcript shows a consistent trend across several families with different genetic backgrounds ($c/c > c/t > t/t$) that is most pronounced at Cr4 (Fig. 3c and Extended Data Fig. 8a). The abundance of the 1A-initiated full transcript, which is in general an order of magnitude less than that of the 1B transcript, does not show a significant difference between genotypes (Fig. 3d and Extended Data Fig. 8b).

The role of *cortex* in wing pattern melanization is not obvious. In *Drosophila*, *cortex* has been primarily associated with meiosis in

ovaries¹¹ (several *cortex* transcripts are expressed in *B. betularia* ovaries and testes; Extended Data Fig. 5). The molecular function of *cortex* is suggested by phylogenetic analysis, which indicates that *Biston cortex* occurs in a lepidopteran sub-group within an insect-specific clade of a protein family containing the cell-cycle regulators *cdc20* and *cdh1*, encoded by *fzy* and *rap* (also known as *fzr*) in *Drosophila* (Extended Data Fig. 9b). These proteins help to regulate fundamental cell division processes such as cytokinesis by presenting substrates to, and activating, the anaphase-promoting complex or cyclosome (APC/C), which ubiquitinates cell-cycle proteins, thereby earmarking them for degradation. Substrate recognition occurs by binding to degrons (short linear motifs

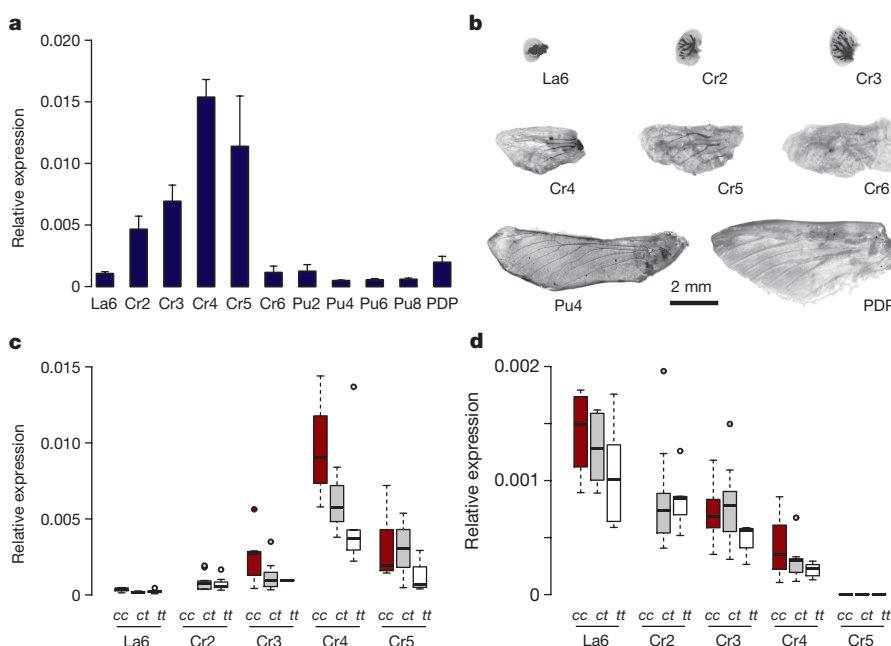


Figure 3 | Relative expression of *cortex* in developing wings of *B. betularia*. **a**, Average expression (across *typica* and *carbonaria* morphs) of all *cortex* splice variants (exons 7–9) relative to the control gene α -*Spec* in wing discs at different developmental stages (La6, sixth instar larvae; Cr2, day 2 crawler; Pu2, day 2 pupae; PDP, post-diapause pupae). Bars are s.e.m. **b**, Scaled images (created from photographs taken by I.J.S.) of *B. betularia* forewings at different stages. **c**, **d**, Tukey plots for relative expression of *cortex* 1B (**c**) and 1A (**d**) full transcripts in developing wings of the three *carbonaria*-locus genotypes (c/c , c/t and t/t) produced within the progeny of a $c/t \times c/t$ cross (no data for c/c at Cr2). Genotypes differ significantly for the 1B full transcript ($P < 0.001$, generalized linear model (GLM)), whereas genotypes do not differ for the 1A full transcript ($P > 0.2$, GLM). (Note the differing y-axis scales.) Equivalent graphs for the progeny of $c/t \times t/t$ crosses (which lack the c/c genotype) are presented in Extended Data Fig. 8.

such as the D box and KEN box). Sequence conservation across lepidopterans and non-lepidopterans reveals a single binding site in cortex (Extended Data Fig. 9c) that probably binds the D box-like¹⁸ degon LXEXXXN¹⁹. This degon binding capability is predicted for both of the full isoforms (1A (441 amino acids) and 1B (407 amino acids), although 1B apparently lacks the N-terminal C box that is usually required for APC/C binding) but not for the alternative isoforms (Extended Data Table 1). These data demonstrate orthology and are consistent with shared function of cortex between *D. melanogaster* and *B. betularia*, although the molecular connection between cell-cycle protein degradation at the APC/C and melanization remains to be determined.

Our results suggest that the *carb*-TE influences adult melanization pattern by increasing the abundance of cortex, perhaps by altering the course of scale-cell heterochrony, with dominance arising through a threshold effect (the 1B full transcript is more abundant in *c/c* than *c/t*). How the *carb*-TE promotes cortex expression is unknown but the general mechanism is predicted to allow the production of *insularia* morphs that are putatively controlled by different mutations within cortex. In combination with parallel findings in *Heliconius* butterflies²⁰, our results support the idea that cortex is a conserved developmental node for generating colour pattern variation in evolutionarily diverse Lepidoptera. However, cortex may not be the only gene in this region involved in patterning, as suggested by recent work on the *B. mori* mutant *Black moth*, which has a similar phenotype to *B. betularia carbonaria*²¹, although none of the genes implicated is differentially expressed among *carbonaria* and *typica* wing discs.

The *carb*-TE is a spectacular example of an adaptively advantageous transposon^{22–24}; its discovery fills a fundamental gap in the peppered moth story and furthers our appreciation of the mechanism underpinning rapid adaptation. A consensus on the general importance of transposable elements for adaptive evolution has yet to emerge^{3,25}. Over longer time frames, phenotypic effects of transposable elements may be obscured by imprecise excision that leaves a minimal trace of the transposable element while retaining the mutant (adaptive) phenotype²⁶. By contrast, we have shown that the *carb*-TE is young, approximately 200 years (generations) old, during which time it has gone from a single mutation to near fixation (regionally) to near extinction—driven by a pulse of environmental change.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 3 July 2015; accepted 22 March 2016.

1. Cook, L. M. The rise and fall of the *carbonaria* form of the peppered moth. *Q. Rev. Biol.* **78**, 399–417 (2003).
2. van't Hof, A. E., Edmonds, N., Dalíková, M., Marec, F. & Saccheri, I. J. Industrial melanism in British peppered moths has a singular and recent mutational origin. *Science* **332**, 958–960 (2011).
3. Brookfield, J. F. Y. Evolutionary genetics: mobile DNAs as sources of adaptive change? *Curr. Biol.* **14**, R344–R345 (2004).
4. Barrett, R. D. H. & Hoekstra, H. E. Molecular spandrels: tests of adaptation at the genetic level. *Nature Rev. Genet.* **12**, 767–780 (2011).
5. Nadeau, N. J. & Jiggins, C. D. A golden age for evolutionary genetics? Genomic studies of adaptation in natural populations. *Trends Genet.* **26**, 484–492 (2010).
6. Martin, A. & Orgogozo, V. The loci of repeated evolution: a catalog of genetic hotspots of phenotypic variation. *Evolution* **67**, 1235–1250 (2013).
7. Stern, D. L. The genetic causes of convergent evolution. *Nature Rev. Genet.* **14**, 751–764 (2013).
8. Savolainen, O., Lascoux, M. & Merilä, J. Ecological genomics of local adaptation. *Nature Rev. Genet.* **14**, 807–820 (2013).
9. Hoekstra, H. E. & Coyne, J. A. The locus of evolution: evo devo and the genetics of adaptation. *Evolution* **61**, 995–1016 (2007).

10. Cook, L. M. & Saccheri, I. J. The peppered moth and industrial melanism: evolution of a natural selection case study. *Heredity* **110**, 207–212 (2013).
11. Chu, T., Henrion, G., Haegeli, V. & Strickland, S. *Cortex*, a *Drosophila* gene required to complete oocyte meiosis, is a member of the Cdc20/fizzy protein family. *Genesis* **29**, 141–152 (2001).
12. Saccheri, I. J., Rousset, F., Watts, P. C., Brakefield, P. M. & Cook, L. M. Selection and gene flow along a diminishing cline of melanized peppered moths. *Proc. Natl Acad. Sci. USA* **105**, 16212–16217 (2008).
13. Clarke, C. A. *Biston betularia*, obligate f. *insularia* indistinguishable from f. *carbonaria* (Geometridae). *J. Lepid. Soc.* **33**, 60–64 (1979).
14. Lees, D. R. & Creed, E. R. Genetics of *insularia* forms of peppered moth, *Biston betularia*. *Heredity* **39**, 67–73 (1977).
15. Kim, Y. & Nielsen, R. Linkage disequilibrium as a signature of selective sweeps. *Genetics* **167**, 1513–1524 (2004).
16. Cook, L. M., Sutton, S. L. & Crawford, T. J. Melanic moth frequencies in Yorkshire, an old English industrial hot spot. *J. Hered.* **96**, 522–528 (2005).
17. Feschotte, C. Transposable elements and the evolution of regulatory networks. *Nature Rev. Genet.* **9**, 397–405 (2008).
18. He, J. *et al.* Insights into degon recognition by APC/C coactivators from the structure of an Acm1-Cdh1 complex. *Mol. Cell* **50**, 649–660 (2013).
19. Whitfield, Z. J., Chisholm, J., Hawley, R. S. & Orr-Weaver, T. L. A meiosis-specific form of the APC/C promotes the oocyte-to-embryo transition by decreasing levels of the polo kinase inhibitor matronomy. *PLoS Biol.* **11**, e1001648 (2013).
20. Nadeau, N. J. *et al.* The gene cortex controls mimicry and crypsis in butterflies and moths. *Nature* <http://dx.doi.org/10.1038/nature17961> (this issue).
21. Ito, K. *et al.* Mapping and recombination analysis of two moth colour mutations, *Black moth* and *Wild wing spot*, in the silkworm *Bombyx mori*. *Heredity* **116**, 52–59 (2016).
22. González, J., Karasov, T. L., Messer, P. W. & Petrov, D. A. Genome-wide patterns of adaptation to temperate environments associated with transposable elements in *Drosophila*. *PLoS Genet.* **6**, e1000905 (2010).
23. Schlenke, T. A. & Begun, D. J. Strong selective sweep associated with a transposon insertion in *Drosophila simulans*. *Proc. Natl Acad. Sci. USA* **101**, 1626–1631 (2004).
24. Schrader, L. *et al.* Transposable element islands facilitate adaptation to novel environments in an invasive species. *Nature Commun.* **5**, 5495 (2014).
25. Casacuberta, E. & González, J. The impact of transposable elements in environmental adaptation. *Mol. Ecol.* **22**, 1503–1517 (2013).
26. Koga, A., Iida, A., Hori, H., Shimada, A. & Shima, A. Vertebrate DNA transposon as a natural mutator: the medaka fish *Tol2* element contributes to genetic variation without recognizable traces. *Mol. Biol. Evol.* **23**, 1414–1419 (2006).

Supplementary Information is available in the online version of the paper.

Acknowledgements The University of Liverpool Centre for Genomic Research (M. Hughes, C. Bourne, R. Eccles, C. Hertz-Fowler and J. Kenny) performed next-generation sequencing and Fragment Analyzer measurements. L. Cook directed us to historical data sources. C. Bergman advised on transposon detection. Population genetics simulations were performed on the University of Liverpool Advanced Research Computing Condor service. This work was supported by Natural Environment Research Council grants NE/H024352/1 and NE/J022993/1.

Author Contributions I.J.S., A.E.v.H. and P.C. designed the study and wrote the paper; P.C., A.E.v.H. and D.J.R. produced the figures; A.E.v.H. directed molecular biology experiments; A.E.v.H., C.J.Y. and J.L. conducted molecular biology experiments; A.E.v.H. constructed the BAC and fosmid tilepaths; A.E.v.H. and A.C.D. assembled, finished and annotated sequences; P.C. analysed population genetic and gene expression data; I.J.S. collected the wild sample; I.J.S. and C.J.Y. reared the samples and performed dissections; D.J.R. and A.E.v.H. built the cortex tree; D.J.R. modelled the cortex structure; M.A.Q. constructed the fosmid library; and A.C.D. and N.H. advised on the design of sequencing strategies.

Author Information The *typica* 1 haplotype (*b-d* interval) reference sequence has been deposited in GenBank under accession number KT182637; The *B. betularia* whole genome sequence has been deposited in the NCBI SRA database under accession number SRX1060178; the cortex splice variants have been deposited in GenBank under accession numbers KT235895–KT235906; *Rps3A* has been deposited in GenBank under accession number JF811439; α -spec has been deposited in GenBank under accession number KT182638. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to I.J.S. (saccheri@liverpool.ac.uk).

METHODS

No formal statistical methods were used to predetermine sample sizes. The experiments were not randomized, and investigators were not blinded to allocation during experiments and outcome assessment.

Wild samples. Moths used for fine mapping and ageing analysis came from a northwest England–north Wales transect sampled in 2002 (ref. 12), with 12 *carbonaria* and 6 *insularia* specimens additionally collected in 2005–2009.

Reference sequences. An extended BAC tiling path was constructed using mapped *B. betularia* genes²⁷, *B. mori* nscaf2829 (SilkDB) orthologues and BAC-end sequences as probes. Combinatorial PCR using BAC-end sequences and internal gene anchors were used to determine the relative positions of the BACs. Fosmids were used to bridge gaps. A minimal tiling path was sequenced as a 3-kb mate-pair library with Roche 454 GS FLX Titanium. Reads were assembled into contigs using Newbler and manually scaffolded using tiled BAC-end sequences and exon order of genes spanning multiple contigs as anchors. The scaffold covers a 3.6-Mb region spanning the mapped genes *Mhc* to *leucine-rich transmembrane protein (Irt)* (GenBank accession numbers HM449891 and HM449887, respectively) with the *carbonaria* polymorphism located towards the centre. A recombination rate estimate within this region of 2.9 cM per Mb was obtained from a total of 350 offspring in 8 crosses screened for recombination between the ends of the 3.6-Mb interval. Three *typica* and one *carbonaria* haplotype sequences were reconstructed using BACs and fosmids for the region spanning locus *b–d* (Fig. 1a). Clones were assigned to haplotypes on the basis of co-segregation of genotypes and phenotypes between parents and sibs of the heterozygous (*carbonaria–typica*) individuals used to generate the BAC (family 67) and fosmid (family 11) libraries. Small assembly gaps caused by repetitive sections were bridged using long capillary Sanger sequences; fosmid clone 25H14, containing the large repetitive transposable element, was sequenced using Pacific Biosystems RS II to 300× coverage using P4-C2 chemistry and assembled using HGAP v2 (Pacific Biosystems). Homopolymer length variation often caused by 454 errors rather than true polymorphisms was verified by Sanger sequencing.

A draft genome assembly was generated from an individual homozygous for the *carbonaria* region. Full-sib *carbonaria–typica* heterozygotes were crossed (family 135) to produce homozygous *carbonaria* offspring, as well as heterozygotes and homozygous *typica*. The *carbonaria* homozygotes were identified using alleles closely linked to the *carbonaria* locus, with more distant loci on either side used to ensure that the haplotype had not been disrupted by recombination. DNA was prepared by phenol–chloroform extraction from a final instar male larva with the gut removed. The genome was sequenced to ~3.5 × coverage on a 454 FLX+ platform and a draft assembly constructed using Newbler. The genome assembly was used for polymorphism discovery, and for tiling path construction using homology to *B. mori*. Single read coverage was used to detect repetitive regions, aiding in single-target primer design, and to confirm the repetitive nature of the *carb-TE*.

The gene content of the *b–d* interval was examined by comparing its sequence with GenBank proteins, expressed sequence tags (ESTs), transcriptomes, and annotated genes in the orthologous region in other Lepidoptera. Tblastx against these orthologous regions and Augustus²⁸ gene prediction were used to detect potentially overlooked genes. All genes were manually annotated and (except for *vcpl*) confirmed using cDNA. The annotation of 11 genes (not including *cortex*) was also subsequently confirmed against a *B. betularia* transcriptome (GenBank SRX371328) assembled with Trinity²⁹. MicroRNAs were found using miRBase with blastn including hairpin precursors. BLAST (blastn, blastx) searches for *carb-TE*-like sequences were performed on NCBI databases (GenBank nucleotide, protein, EST, transcriptome), independently curated lepidopteran genome assemblies (for example, SilkDB), and RepBase (19.09).

Fine mapping. The interval containing the *carbonaria* polymorphism was narrowed down to a section bordered on both sides by evidence of *carbonaria* haplotype breakdown caused by recombination. Polymorphisms at regular intervals in the *b–d* region (Fig. 1a; Supplementary Table 2) were genotyped in wild-caught *carbonaria*, *typica* and *insularia* (105, 33 and 30 individuals, respectively). We conservatively used only homozygous genotypes to set these boundaries because the dominance of *carbonaria* obscures the assignment of alleles in heterozygous genotypes to a certain morph haplotype. The four contiguous haplotype sequences (one *carbonaria* and three *typica*) constructed from BACs and fosmids were aligned between these narrowed-down boundaries and examined for polymorphisms that were distinct in the *carbonaria* haplotype relative to all three *typica* haplotypes, resulting in 87 *carbonaria* candidate polymorphisms (Extended Data Fig. 1 and Supplementary Table 1). With the exception of *carbonaria_candidate_45*, wild-caught *typica* were genotyped at all loci by means of PCR–restriction-fragment length polymorphism (RFLP), PCR–indel or sequencing. Depending on the frequency of the candidate alleles in the *typica* sample, 16 to 283 *typica* (32 to 566 *typica* haplotypes) were used for exclusion. *Carbonaria_candidate_25* was present

in only one out of 566 *typica* haplotypes. The *typica* phenotype of this individual (12-2002-01) was confirmed, as was the presence of the *carbonaria_candidate_25* allele from independently extracted DNA. A very large indel, later identified as the true *carbonaria* polymorphism (*carbonaria_candidate_45*), that could not be bridged by PCR required an alternative present/absent screening approach which also provided a positive control for absence haplotypes (to distinguish insert absence from PCR failure). A three-primer PCR was designed with two primers flanking the indel and a third within the insert, relatively close to the indel boundary (Extended Data Fig. 2). The assay was validated using a family known to include all three genotypes (family 135, Extended Data Fig. 2).

Inferring haplotypes and the age of the *carbonaria* mutation. A set of 177 individuals, including 105 *carbonaria* individuals, was genotyped at 119 polymorphic loci within 28 PCR products, stretching across ~400 kb (Supplementary Table 2). *Carbonaria* haplotypes were inferred using SHAPEIT³⁰ and the position (interval) of recombination breakpoints inferred based on two or more consecutive phase-switched polymorphisms. High repeatability of the phasing outcomes was verified by resampling, and switch errors were minimized by including known haplotypes and classifying only two types (melanic and non-melanic). Indices of multilocus linkage disequilibrium (r_d) were calculated from polymorphisms within each PCR fragment and the *carbonaria* locus across the 400-kb interval³¹. Their significance was assessed using 999 Monte-Carlo permutations. The pattern of introgression of the *carbonaria* haplotype into background haplotypes (that is, *typica* and *insularia* morph alleles) was assessed using ChromoPainter v2 (ref. 32) to search for contiguous blocks that match the *carbonaria* haplotype, thus generating the ‘expectation painting’ of background haplotypes.

The age of the *carbonaria* mutation was inferred with a simulation-based approach. The analysis was performed in three steps. First, 1,000,000 time-forward trajectories of the *carbonaria* phenotype were sampled, using a Metropolis–Hastings algorithm, depending on their likelihood given historical phenotypic frequencies (Supplementary Table 3), and conditional to their starting date (x_0) and population size (N). Second, recombination patterns were simulated using the sampled trajectories, in populations of size N , and a fixed recombination rate of 2.9 cM per Mb (males only). This process yielded sample distributions of the closest recombination breakpoint relative to the *carbonaria* locus. Finally, the likelihood of the simulated distributions given the empirical recombination pattern was computed and averaged across simulations to estimate the probability density of the mutation age (x_0). For full details, see Supplementary Methods.

Code availability. Code available on request.

Expression and alternative transcripts of *cortex*. Offspring from either heterozygous *carbonaria/typica* (*c/t*) × homozygous *t/t* crosses segregating 1:1 or *c/t* × *c/t* crosses segregating 1 *c/c*: 2 *c/t*: 1 *t/t* were used for end-point reverse transcription PCR (RT–PCR) and real-time quantitative PCR (qPCR) experiments. Caterpillars were reared on grey willow (*Salix cinerea*). Wing discs (forewings and hindwings) were dissected from final (sixth) instar larvae, crawlers or prepupae (days 2–6 from the start of crawling stage), pre-diapause pupae (days 2–8 from pupation, at which point they have entered diapause) and post-diapause pupae (wing discs staged into six categories), and stored in RNAlater (Ambion). RNA was extracted with TRIzol and cDNA synthesized with SuperScript III (Invitrogen)–oligo(dT). The genotype–phenotype (adult morph) of each wing disc specimen was determined with the *carb-TE* three primer PCR (and verified by sequencing a linked single nucleotide polymorphism (SNP), *carbonaria_candidate_25*). Relative abundance and qPCR data were analysed using generalized linear (mixed) models (GLM). See Extended Data Fig. 8c for sample sizes.

Quantitative PCR experiments were designed to measure the relative abundance of *cortex* transcripts, either of all transcripts combined (using primers in exons 7 and 9) or full transcripts only (primers in exons 1A–3 and 1B–3, as exon 3 is effectively exclusive to the full transcripts (Extended Data Fig. 7)). DNase treatment was not performed, but for exons 7–9 qPCR co-amplification of genomic DNA was prevented by positioning the reverse primer on the exon 8–9 boundary (this was not a concern for exons 1–3 qPCR because the large first intron precluded genomic DNA amplification). We chose 40S ribosomal protein *S3a* (*RpS3a*)³³ and α -spec³⁴ as two single-copy autosomal housekeeping genes. Primer sequences are listed in Supplementary Table 4. Annealing temperatures were optimised to 66 °C and amplicons were confirmed to produce single bands on agarose gels. cDNA was diluted 1:1 with water to allow template volumes within the accuracy range of the pipette used. Quantitative PCRs for target and control were run in three replicates using Kapa SYBR Fast qPCR Universal under recommended conditions on a Roche LightCycler 480 with 45 cycles and a melting curve. As both control genes gave similar results, only α -spec was used for the entire sample.

Alternative transcription starts of *cortex* were searched for using 5′ rapid amplification of cDNA ends (RACE) on RNA extracted from 15 wing disc samples covering a wide range of stages and *c/c*, *c/t* and *t/t* genotypes, and also from

whole pupae and testes. *Cortex*-specific cDNA was synthesized with SuperScript III and a gene-specific negative strand primer; 5' cytosine extension was added using terminal transferase (NEB) and deoxycytidine triphosphates (dCTPs). The single-stranded cDNA was made double-stranded and a target sequence for amplification incorporated in a single extension cycle (LongAmp Hot Start, NEB) with an oligonucleotide containing a 5' primer recognition site and a 3' poly-G tail. PCR was performed using a forward primer matching the synthetic 5' end and a nested *cortex*-specific reverse primer. The amplicons were sequenced using a second nested primer. The alternative first exons were confirmed by Sanger sequencing with forward primers inside the newfound exons to generate clean sequence without the background noise commonly observed with 5' RACE.

The complete pattern of *cortex* splice variation was examined with end-point RT-PCR using primers Bb_cort_exon1A_F or Bb_cort_exon1B_F and Bb_cort_exon9_R (for primer sequences see Supplementary Table 4). PCR conditions were 60 °C annealing, 40 cycles, 75 s extension, 25 µl total volume, 3 µl wing disc cDNA, LongAmp Taq DNA polymerase (NEB). A Fragment Analyzer (Advanced Analytical) was used to estimate the size and relative abundance of amplicons within each individual, after normalizing samples to a concentration range of ~1–10 ng µl⁻¹. The concentration of each fragment peak was calculated using PROSize (Advanced Analytical), and the relative abundance was computed as the concentration of a splice variant divided by the sum of all fragment concentrations within that individual profile. The *cortex* splice variant amplicons were sequenced as two pools (*t/t* and *c/t*) using Pacific Biosystems RS II with P6-C4 chemistry and the insert reads extracted using smrtportal (Pacific Biosystems). Reads that contained exon 1A or 1B and exon 9 were used to validate the sequence composition and relative abundance of spliced gene isoforms.

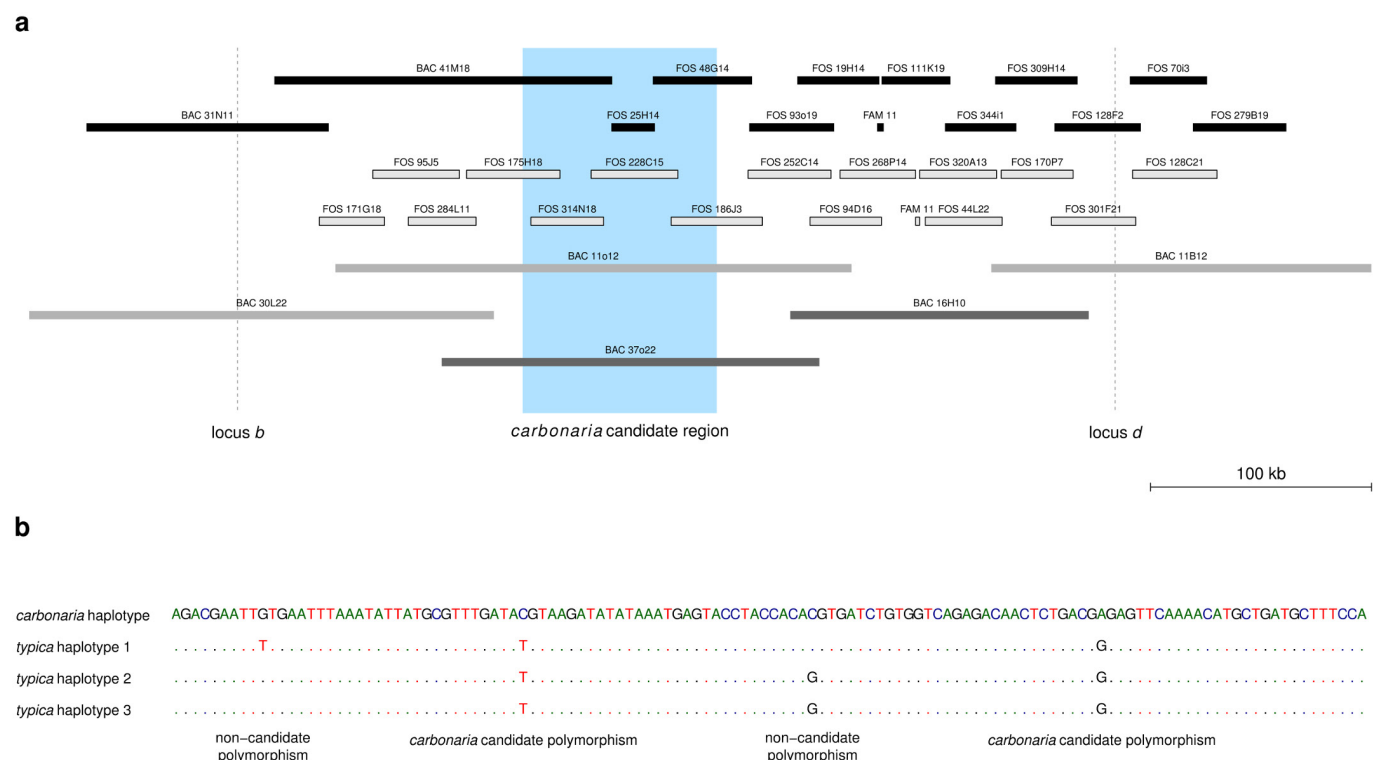
No part of *carb*-TE was detected in *cortex* transcripts, either with PacBio sequencing or with PCR using various primer combinations where one primer lies within the transposon and the other matches a *cortex* exon. However, a *carb*-TE-like partial sequence was amplified (with primers within repeat units) from both *typica* and *carbonaria* morph cDNA synthesized using *carb*-TE primers, implying that these RNA sequences are transcribed from non-allelic homologues of the *carb*-TE.

Expression of alternative candidate genes. Two *B. mori* adult melanism/patterning mutations, *Black moth* (*Bm*) and *Wild wing spot* (*Ws*), were recently mapped to a region partially orthologous to the *carbonaria* interval²¹. In this study, end-point PCR showed complete absence of *cortex* expression in pupal stages and adults but potentially important prepupal stages were not examined. Three neighbouring genes (*BGIBMGA005658*, *BGIBMGA005657* and *BGIBMGA005655*) did show convincing differences between the wild type and both mutants even though these genes lie outside the *Ws* mapping interval. We performed equivalent end-point RT-PCR for the three orthologues in *B. betularia* to determine whether morph-expression associations existed between *carbonaria* and *typica* (comparing *c/t* and *t/t* genotypes for wing disc stages Cr4, Cr6, Pu2, Pu4 and PDP). PCR

conditions were as for *cortex* 1A/1B–9 end-point PCRs, except for 45-s extension (for primer sequences, see Supplementary Table 4).

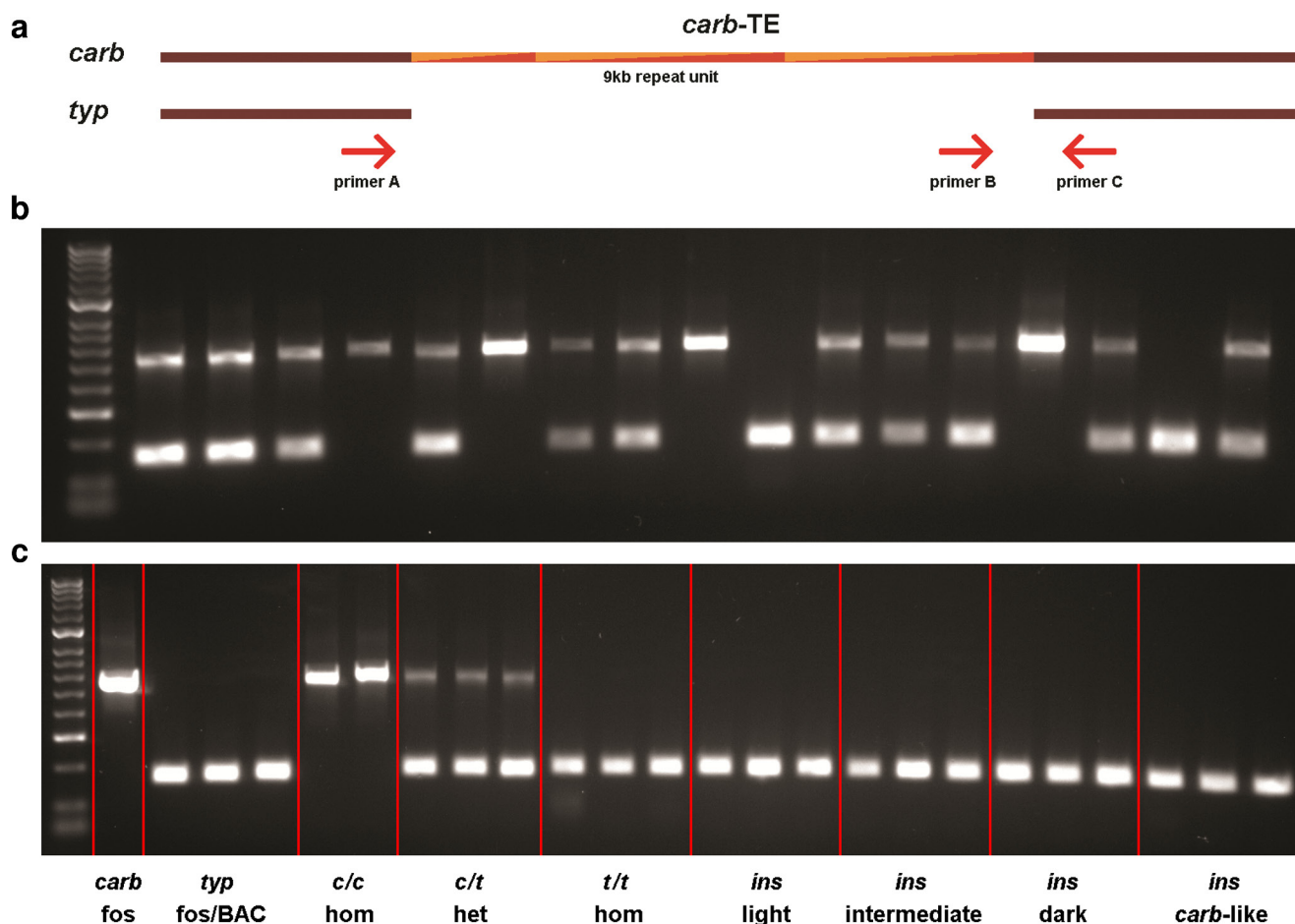
Cortex phylogeny and protein modelling. Cortex sequences derived from database searches (Supplementary Table 5) were supplemented with a selection of *chd1* and *cdc20/fzy* sequences from model organisms and the set aligned with MAFFT³⁵ (Supplementary Data 2). The central propeller domain was isolated and used for bootstrapped phylogenetic analysis with MEGA 6 (ref. 36) employing its Maximum Likelihood algorithm and the JTT matrix-based model. Any gapped positions were ignored. Homology models of *B. betularia* and *D. melanogaster* cortex proteins were made with MODELLER³⁷ and Consurf³⁸ was used to map protein sequence conservation to their respective surfaces among lepidopteran or non-lepidopteran cortex proteins.

27. van't Hof, A. E. *et al.* Linkage map of the peppered moth, *Biston betularia* (Lepidoptera, Geometridae): a model of industrial melanism. *Heredity* **110**, 283–295 (2013).
28. Stanke, M. & Morgenstern, B. AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. *Nucleic Acids Res.* **33**, W465–W467 (2005).
29. Grabherr, M. G. *et al.* Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnol.* **29**, 644–652 (2011).
30. Delaneau, O., Marchini, J. & Zagury, J.-F. A linear complexity phasing method for thousands of genomes. *Nature Methods* **9**, 179–181 (2011).
31. Agapow, P.-M. & Burt, A. Indices of multilocus linkage disequilibrium. *Mol. Ecol. Notes* **1**, 101–102 (2001).
32. Lawson, D. J., Hellenthal, G., Myers, S. & Falush, D. Inference of population structure using dense haplotype data. *PLoS Genet.* **8**, e1002453 (2012).
33. Baxter, S. W. *et al.* Genomic hotspots for adaptation: the population genetics of Müllerian mimicry in the *Heliconius melpomene* clade. *PLoS Genet.* **6**, e1000794 (2010).
34. Reed, R. D., McMillan, W. O. & Nagy, L. M. Gene expression underlying adaptive variation in *Heliconius* wing patterns: non-modular regulation of overlapping cinnabar and vermilion prepatterning. *Proc. R. Soc. Lond. B* **275**, 37–45 (2008).
35. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).
36. Tamura, K., Stecher, G., Peterson, D., Filipi, A. & Kumar, S. MEGA6: molecular evolutionary genetics analysis version 6.0. *Mol. Biol. Evol.* **30**, 2725–2729 (2013).
37. Šali, A. & Blundell, T. L. Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol.* **234**, 779–815 (1993).
38. Ashkenazy, H., Erez, E., Martz, E., Pupko, T. & Ben-Tal, N. ConSurf 2010: calculating evolutionary conservation in sequence and structure of proteins and nucleic acids. *Nucleic Acids Res.* **38**, W529–W533 (2010).
39. Muñoz-López, M. & García-Pérez, J. L. DNA transposons: nature and applications in genomics. *Curr. Genomics* **11**, 115–128 (2010).
40. Chang, L., Zhang, Z., Yang, J., McLaughlin, S. H. & Barford, D. Atomic structure of the APC/C and its mechanism of protein ubiquitination. *Nature* **522**, 450–454 (2015).



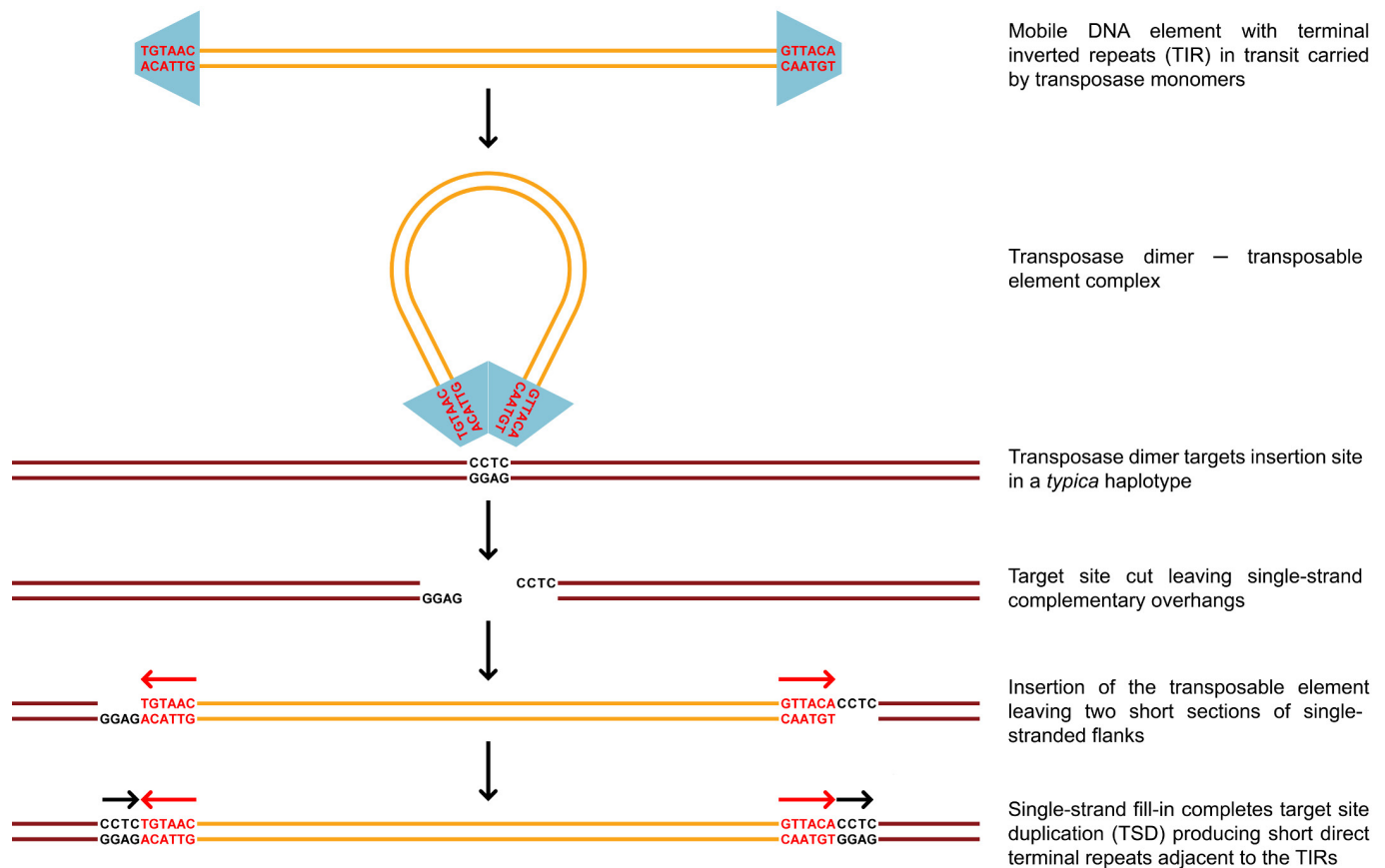
Extended Data Figure 1 | BAC and fosmid haplotype tilepaths used to define *carbonaria* candidate polymorphisms. **a**, BAC and fosmid tilepaths of the *carbonaria* haplotype (black bars) and three *typica* haplotypes (different shades of grey). Two small regions not covered by BACs or fosmids were reconstructed using parent and offspring sequences from the same heterozygous family (FAM11). The positions of loci *b* and *d* (see Figure 1) are indicated by the dashed lines, and the *carbonaria* candidate region is highlighted blue. Fosmid 25H14 containing *carb*-TE

appears small because it is aligned against the *typica* reference sequence, which does not include the *carb*-TE. **b**, Alignment of three *typica* haplotypes against the *carbonaria* haplotype for a short section within the *carbonaria* candidate region, showing SNPs (dots are nucleotides identical to the *carbonaria* sequence). Polymorphisms in which all three *typica* alleles differed from *carbonaria* were treated as *carbonaria* candidates; polymorphisms in which the same allele occurred in *carbonaria* and at least one *typica* were excluded from further consideration.



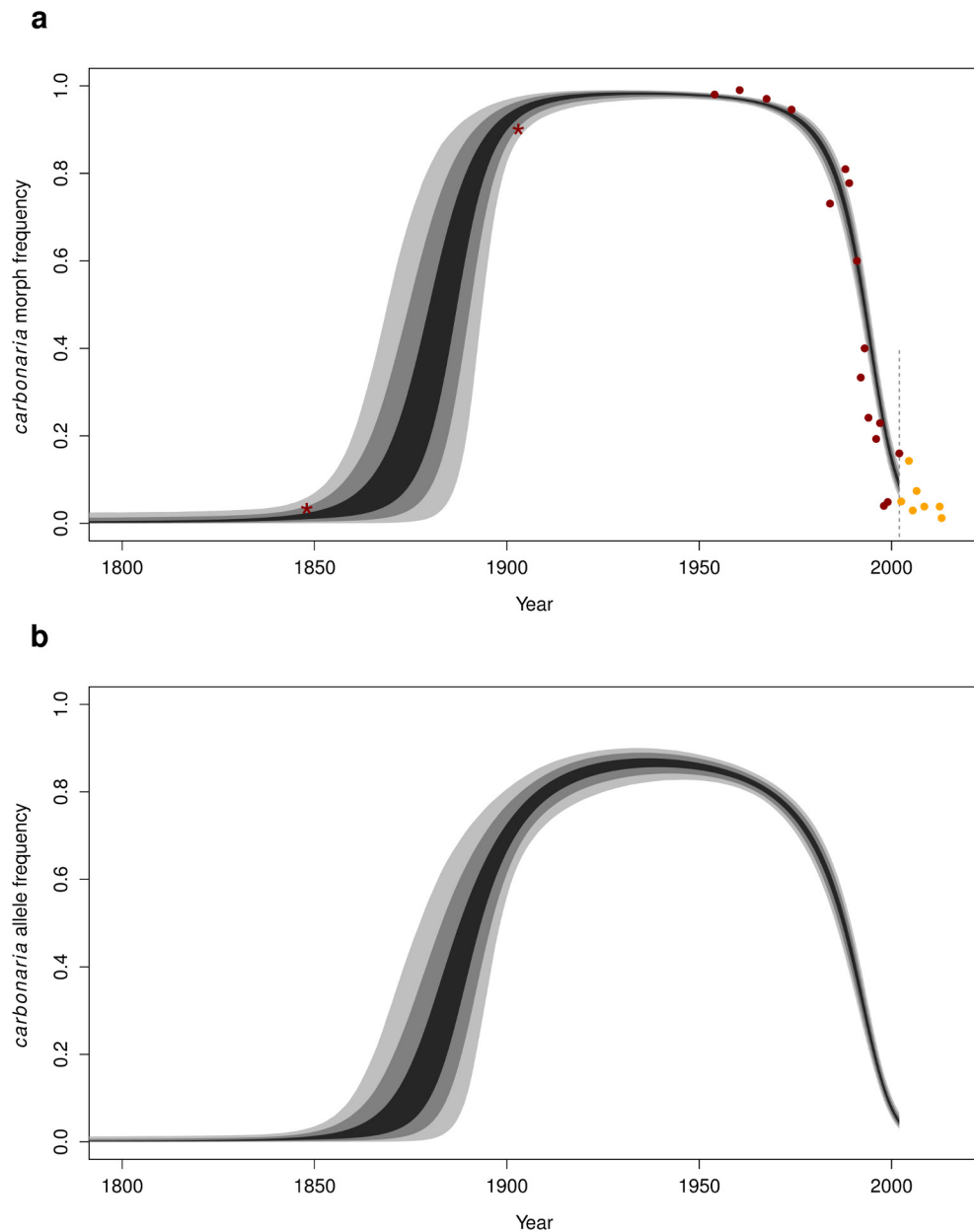
Extended Data Figure 2 | Validation of the 3-primer PCR *carb*-TE genotyping assay in a family and its application in a variety of wild-caught moths. a, Schematic alignment of *carbonaria* and *typica* haplotypes showing the position of the three primers (A, B and C, not to scale) used in the same PCR to detect the presence and absence of the 22 kb *carb*-TE. In the presence of the *carb*-TE, primers A and C are too far apart to generate a product; the repeat structure of the *carb*-TE presents three annealing sites for primer B but only the shortest primer B–C combination is amplified when using 45-s extension (primer sequences are listed in Supplementary Table 1). **b**, *carb*-TE genotypes for father (lane 2), mother (lane 3) and 15 offspring (lanes 4–18); the two brightest bands in the size ladder are 300 bp and 1 kb (lane 1). The parents were full siblings and

known to be heterozygous (*c/t*), and therefore expected to generate *c/c*, *c/t* and *t/t* offspring. The larger band (primers B–C) indicates the presence of the *carb*-TE and the smaller band (primers A–C) its absence (*typica* allele in this family); heterozygotes have both bands. The individual in lane 15 (135F1-12) is the homozygous male used for whole genome sequencing. **c**, Presence or absence of the *carb*-TE in a *carbonaria* haplotype fosmid clone (lane 2), three different *typica* haplotype clones (lanes 3–5; one fosmid, two BACs), wild *carbonaria* homozygotes (lanes 6 and 7), wild *carbonaria* heterozygotes (lanes 8–10), *typica* with a flanking haplotype similar to the *carbonaria* haplotype but lacking the *carb*-TE (lanes 11–13), light *insularia* (lanes 14–16), intermediate *insularia* (lanes 17–19), dark *insularia* (lanes 20–22) and *carbonaria*-like *insularia* (lanes 23–25).



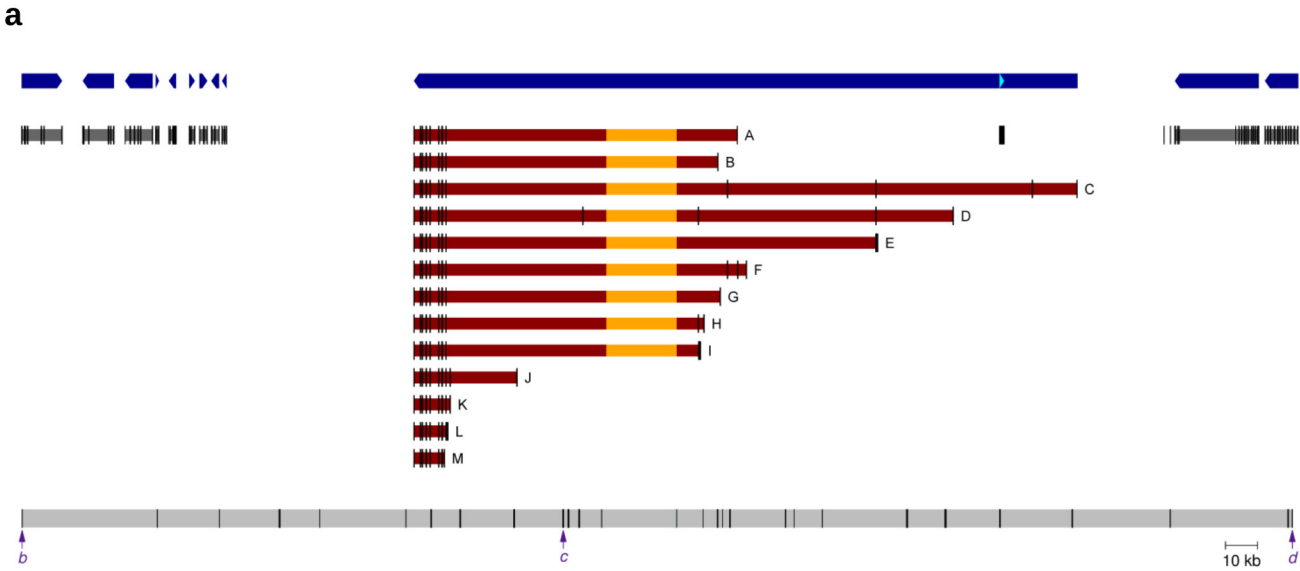
Extended Data Figure 3 | Hypothetical reconstruction of the birth of the *carbonaria* allele. Class II non-autonomous DNA transposition is mediated by two transposase monomers linked to terminal inverted repeats (TIR). The monomers form a dimer at the target site that is cleaved to leave short direct repeated overhangs. The transposable element including TIRs is inserted and finally the single-stranded cleaved sites are

filled in to complete the target site duplication³⁹. The unduplicated target site motif (CCTC) is common, possibly ubiquitous, in all non-*carbonaria* (*typica* and *insularia*) haplotypes, but a *typica* ancestor is more likely given the pattern of haplotype similarities and the presumed prevalence of *typica* haplotypes around 1800.



Extended Data Figure 4 | The rise and fall of *carbonaria* in the Manchester area. a, Frequency of the *carbonaria* phenotype from ~1800 to 2009. **b,** Corresponding frequencies of the *carbonaria* allele. The envelopes show the confidence intervals (50%, 90% and 99%) for the simulated trajectories. Dark-red dots, observations falling within the simulated

trajectories; orange dots, additional data collected after 2002 (year during which >85% of the field sample was collected). Stars indicate likely frequencies where historical data are scarce. Data and sources are listed in Supplementary Table 3.

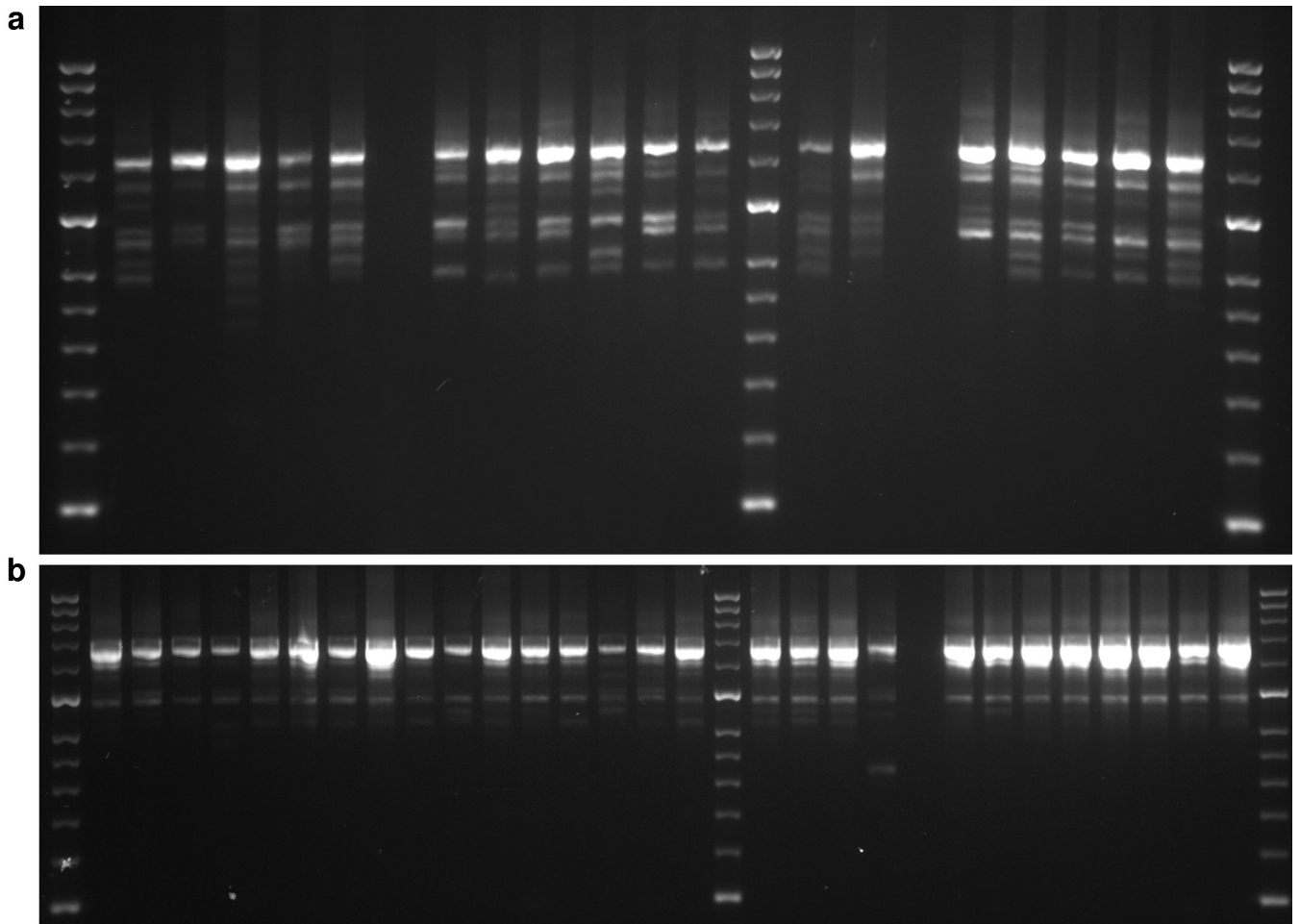


b

Origin	wd/pup	wd	testes	wd	wd	testes	wd	testes	wd	wd	testes	wd	testes
Exon 1	A	B	C	D	E	F	G	H	I	J	K	L	M
testes	-	++	+	-	+	+	-	-	-	-	++	-	-
ovaries	-	+	-	-	-	-	-	-	-	-	++	+	-
La6_tt	+	-	-	-	+	-	-	-	-	-	-	-	-
La6_cc	+	-	-	-	+	-	-	-	-	-	-	-	-
Cr2_tt	+	++	-	+	-	-	-	-	-	-	-	-	-
Cr2_ct	+	++	-	-	-	-	-	-	-	-	+	-	-
Cr4_tt	-	+++	-	-	-	-	-	-	-	-	-	+	+
Cr4_cc	-	+++	-	-	-	-	-	-	-	-	-	+	-
Cr6_tt	-	+	+	-	-	-	-	-	-	-	+	-	-
Cr6_cc	-	+	+	-	-	-	-	-	-	-	+	-	-
Pu2_tt	+	+	-	-	+	-	-	-	-	-	-	+	-
Pu2_ct	+	+	-	-	-	-	-	-	-	-	-	+	-
PDP_tt	+	++	-	-	-	-	-	-	-	-	-	-	-
PDP_ct	+	++	-	-	-	-	-	-	-	-	-	-	-

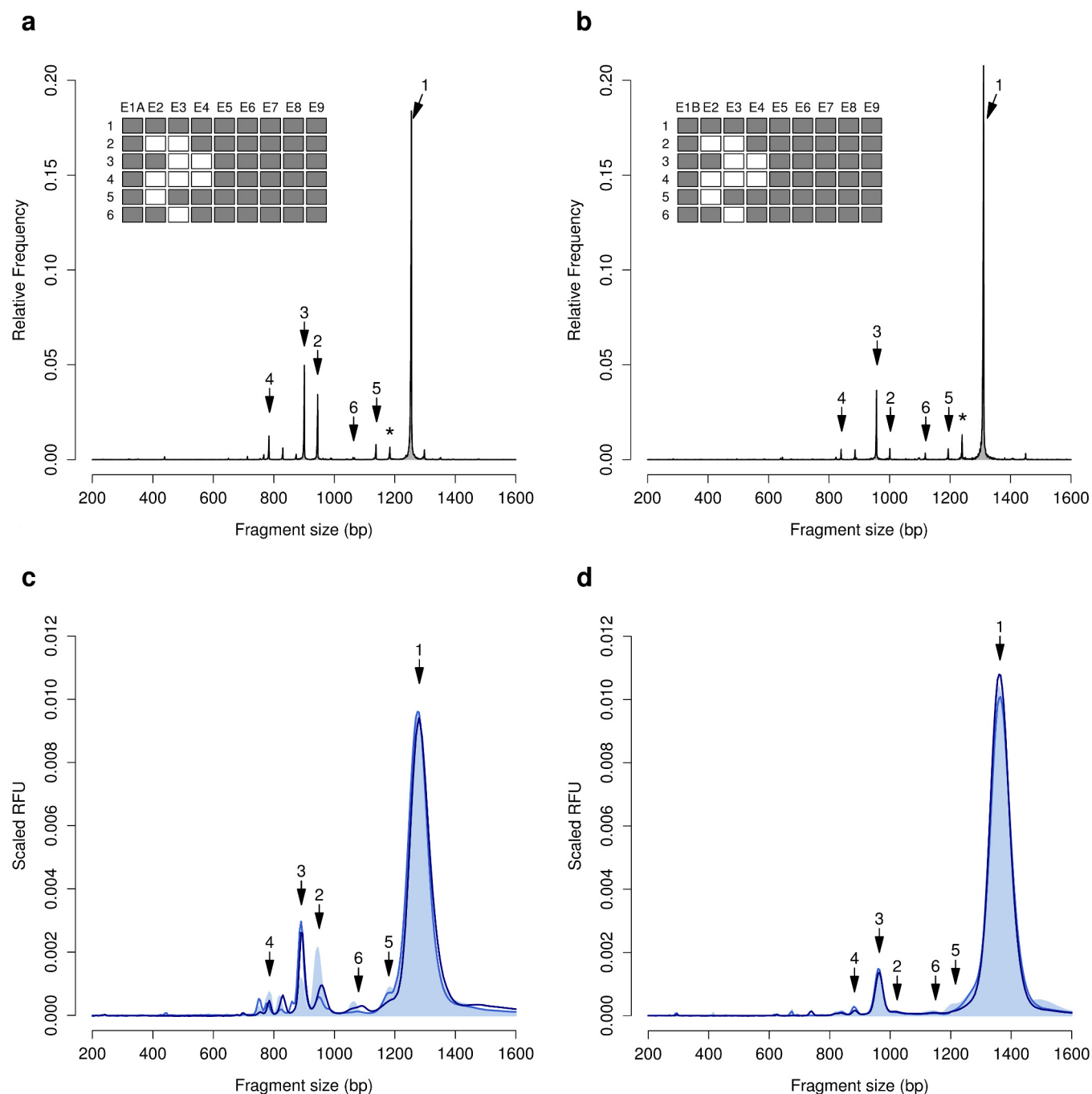
Extended Data Figure 5 | Position and tissue-specific expression of alternative first exons of the gene *cortex*. **a**, Illustration of *cortex* exon structure indicating the positions of thirteen alternative transcription starts and subsequent exons relative to the flanking genes in the *b-d* region (position of *carb*-TE indicated by orange bar). **b**, Expression of different starting position *cortex* transcripts. End-point RT-PCR with reduced cycles (35) was used to exclude transcripts with negligible dosage. Amplicon intensities are scaled between + (faint but visible) and

+++ (strong PCR product). Negative PCRs represent expression below the detection threshold; this may even occur in 'origin' tissue types (wing disc/pupa/testes) in which the alternative starts were discovered owing to the fact that 5' RACE used ~20 times the amount of RNA template relative to the standard cDNA synthesis for the 35 cycle end-point PCRs. Ovaries were not used for 5' RACE, which may have caused gonad expression bias towards testes. Test tissues are sixth instar larvae gonads and wing discs at different developmental stages (abbreviations as in Fig. 3).



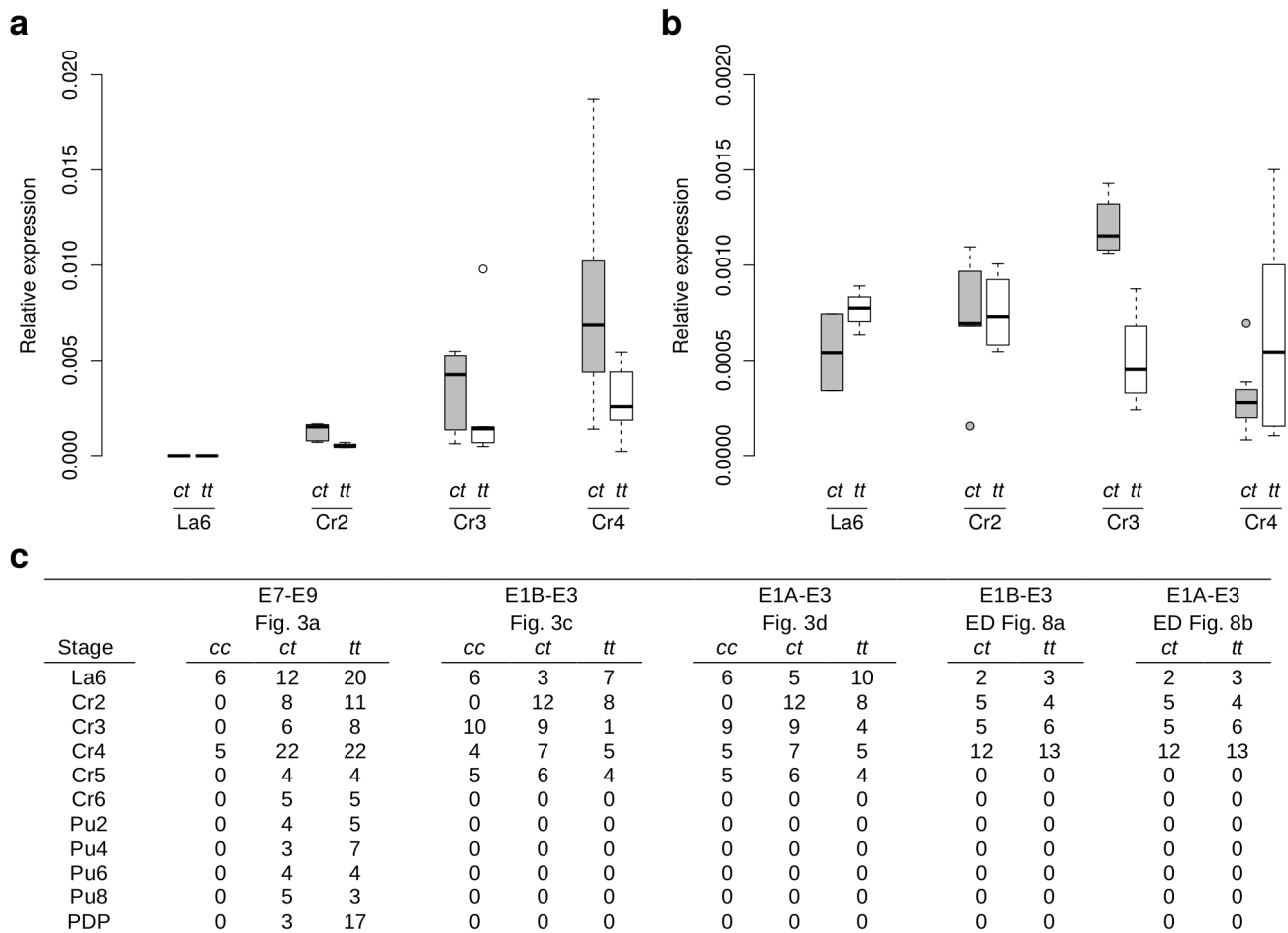
Extended Data Figure 6 | Examples of *cortex* splice variation pattern in *typica* and *carbonaria* developing wing discs. End-point PCR on wing disc cDNA amplified with primers in the first and last exons (E1–E9), with *typica* individuals to the left of the central ladder (the two brightest bands

in the size ladder are 300 bp and 1 kb) and *carbonaria* individuals (all *c/t* heterozygotes) to the right of the central ladder. **a**, Exon 1A variants in Cr2 stage. **b**, Exon 1B variants in Cr4 stage. (See Fig. 3 for stage abbreviations.)



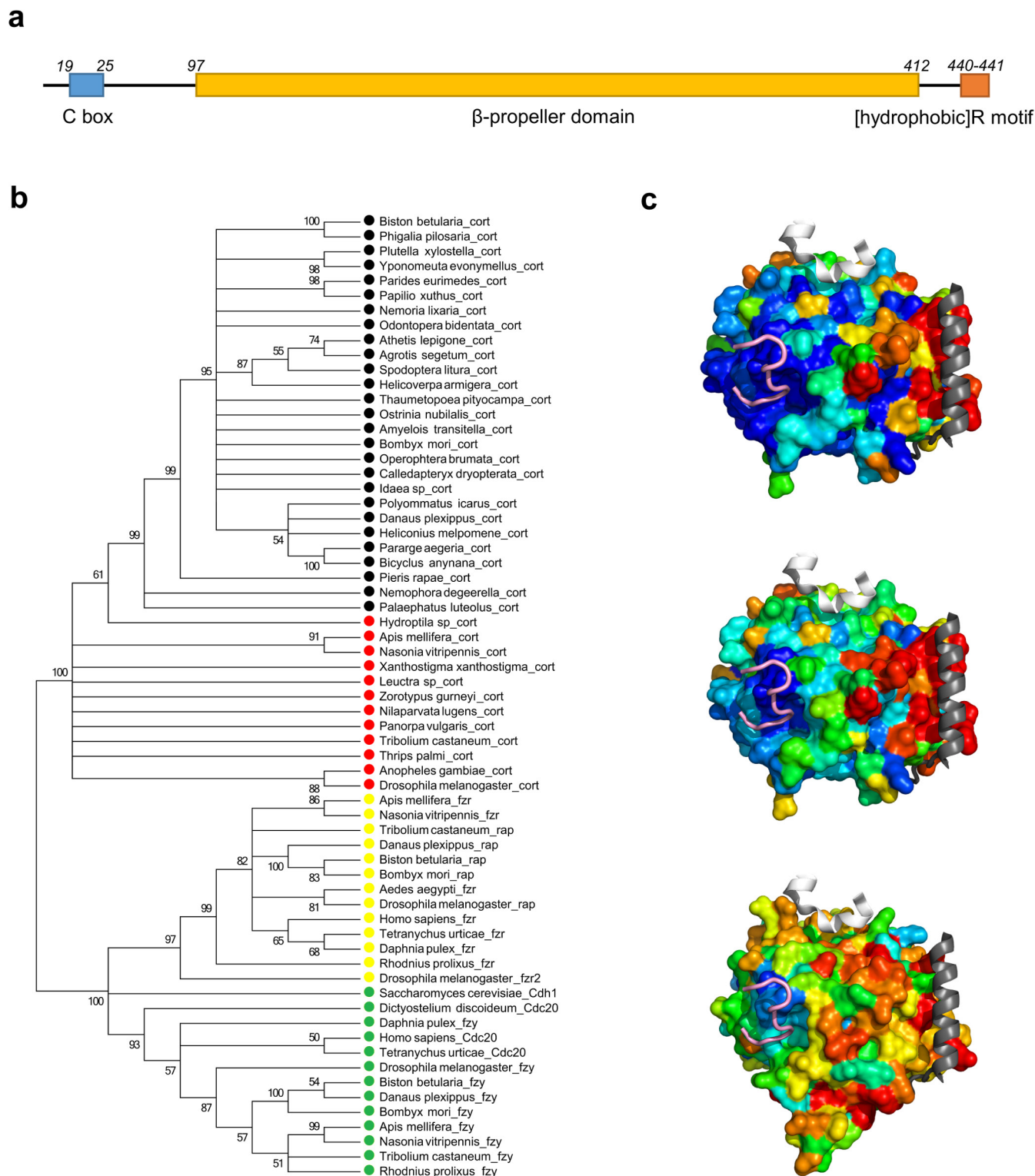
Extended Data Figure 7 | Exonic structure and size distributions of *cortex* splice variants amplified by end-point RT-PCR with primers in exon 1A or 1B and exon 9. Size distributions of the PacBio reads are displayed for the two alternative first exons 1A (**a**) and 1B (**b**) of *cortex*. **c, d**, Comparison of *carbonaria* locus genotypes (*t/t* pale blue fill, *c/t* light blue line, *c/c* dark blue line) measured with Fragment Analyzer. Relative fluorescence units (RFU) were averaged across individuals for fragments amplified with E1A-E9 (**c**) or E1B-E9 (**d**) primers. Prior to averaging, RFUs were standardized so that the total fluorescence (area under the

curve) per individual scaled to 1. Arrows with the same numbers denote either similar exonic structure (E1A versus E1B variants) or fragment identity between the two sources of data (PacBio reads and Fragment Analyzer). Exonic structure of the six main splice variants is represented in matrices (**a, b**), in which white cells represent skipped exons in a splice variant (asterisk indicates full transcript in which the first 71 bp of exon 6 are missing). Apparent differences among melanic and non-melanic for 1A number 2 and number 3 splice variants were not consistent among families.



Extended Data Figure 8 | Tukey plots for relative expression of *cortex* full transcript in developing wing discs. *c/t* heterozygotes are compared with *t/t* homozygotes produced from *c/t* × *t/t* crosses (starting with exon 1B (a) or exon 1A (b)). Genotypes differ significantly for 1B full transcript

($P = 0.001$, GLM), whereas genotypes do not differ for 1A full transcript ($P > 0.5$, GLM). Note the differing y-axis scales. c, Sample sizes for *cortex* qPCR experiments by wing disc developmental stage and *carbonaria*-locus genotype.



Extended Data Figure 9 | Orthology and functional domain conservation of cortex protein. **a**, Schematic illustration, not to scale, of molecular features of *B. betularia* cortex protein sequence. **b**, Bootstrapped Maximum Likelihood consensus tree calculated with MEGA 6 of *fzy/cortex* derived from the propeller domain of the alignment in Supplementary Data 2. Branches are collapsed where partitions were reproduced in less than half of bootstrap replicates. Major groups containing lepidopteran cortex (black circles), non-lepidopteran cortex (red circles), *fzy/rap* (yellow circles) or *fzy/cdc20/cdh1* proteins (green circles) are similarly unequivocally defined in trees obtained by neighbour joining or maximum parsimony methods (not shown). **c**, 3D protein sequence conservation mapping of lepidopteran cortex sequences onto a homology model of *B. betularia* cortex (top); all cortex

sequences onto the same *B. betularia* model (middle); non-lepidopteran cortex sequences onto a model of *D. melanogaster* cortex (bottom). Molecular surfaces are shown in PyMOL using a spectrum from high (blue) to low (red) conservation. The mapping reveals the shared presence of a presumed inter-blade D box-like degron-binding site (pink segment is superimposed D box-mimicking sequence from the structures of human APC/C (PDB accession 4ui9)⁴⁰). In contrast, there is much weaker conservation of surface regions corresponding to facial KEN box or helical specificity determinant sites (white and grey ribbons, respectively, from the same structure), suggesting that cortex proteins lack these functionalities. Note that the greater sequence variability in the non-lepidopteran set leads to lower overall sequence conservation (bottom) but that overall patterns in all panels are similar.

Extended Data Table 1 | Predicted functionality of *B. betularia* cortex isoforms (starting with exon 1A or 1B)

Feature known in Cdh1/Cdc20 (function) and its potential conservation							
Isoform*	Length (residues)	Binding to APC/C			Binding to degrons (see Ext. Dat. Fig. 9)		
		C-box: DRFVVPR (binds Apc8 subunit of APC/C)	Segments 2 & 4 (bind Apc1)	[hydrophobic]R C-terminus (binds Apc3)	Inter-blade recognition site for LxExxxN degron	Facial recognition of KEN-box degron	Recognition of helical specificity determinant
1A	441	✓	✗	✓	✓	✗	✗
1B	407	✗	✗	✓	✓	✗	✗
2A	291	✗	✗	✓	✗ [†]	✗	✗
2B	291	✗	✗	✓	✗ [†]	✗	✗
3A	323	✓	✗	✓	✗	✗	✗
3B	289	✗	✗	✓	✗	✗	✗
4A	284	✗	✗	✓	✗	✗	✗
4B	270	✗	✗	✓	✗	✗	✗
5A	402	✗	✗	✓	✓	✗	✗
5B	270	✗	✗	✓	✗	✗	✗
6A	291	✗	✗	✓	✗ [†]	✗	✗
6B	291	✗	✗	✓	✗ [†]	✗	✗

*Isoforms as defined in Extended Data Fig. 7.

†As the region lost from the propeller fold constitutes approximately a single blade, it is possible that these, and only these, truncated-propeller forms may still fold stably.

The gene *cortex* controls mimicry and crypsis in butterflies and moths

Nicola J. Nadeau^{1,2}, Carolina Pardo-Díaz³, Annabel Whibley^{4,5}, Megan A. Supple^{2,6}, Suzanne V. Saenko⁴, Richard W. R. Wallbank^{2,7}, Grace C. Wu⁸, Luana Maroja⁹, Laura Ferguson¹⁰, Joseph J. Hanly^{2,7}, Heather Hines¹¹, Camilo Salazar³, Richard M. Merrill^{2,7}, Andrea J. Dowling¹², Richard H. ffrench-Constant¹², Violaine Llaurens⁴, Mathieu Joron^{4,13}, W. Owen McMillan² & Chris D. Jiggins^{2,7}

The wing patterns of butterflies and moths (Lepidoptera) are diverse and striking examples of evolutionary diversification by natural selection^{1,2}. Lepidopteran wing colour patterns are a key innovation, consisting of arrays of coloured scales. We still lack a general understanding of how these patterns are controlled and whether this control shows any commonality across the 160,000 moth and 17,000 butterfly species. Here, we use fine-scale mapping with population genomics and gene expression analyses to identify a gene, *cortex*, that regulates pattern switches in multiple species across the mimetic radiation in *Heliconius* butterflies. *cortex* belongs to a fast-evolving subfamily of the otherwise highly conserved fuzzy family of cell-cycle regulators³, suggesting that it probably regulates pigmentation patterning by regulating scale cell development. In parallel with findings in the peppered moth (*Biston betularia*)⁴, our results suggest that this mechanism is common within Lepidoptera and that *cortex* has become a major target for natural selection acting on colour and pattern variation in this group of insects.

In *Heliconius*, there is a major effect locus, *Yb*, that controls a diversity of colour pattern elements across the genus. It is the only locus in *Heliconius* that regulates all scale types and colours, including the diversity of white and yellow pattern elements in the two co-mimics *H. melpomene* and *H. erato*, and whole-wing variation in black, yellow, white, and orange/red elements in *H. numata*^{5–7}. In addition, genetic variation underlying the *Bigeye* wing pattern mutation in *Bicyclus anynana*, melanism in the peppered moth, *Biston betularia*, and melanism and patterning differences in the silkworm, *Bombyx mori*, have all been localized to homologous genomic regions^{8–10} (Fig. 1). Therefore, this genomic region appears to contain one or more genes that act as major regulators of wing pigmentation and patterning across the Lepidoptera.

Previous mapping of this locus in *H. erato*, *H. melpomene* and *H. numata* identified a genomic interval of about 1 Mb (refs 11–13) (Extended Data Table 1), which also overlaps with the 1.4-Mb region containing the *carbonaria* locus in *B. betularia*⁹ and a 100-bp non-coding region containing the *Ws* mutation in *B. mori*¹⁰ (Fig. 1). We used a population genomics approach to identify the single nucleotide polymorphisms (SNPs) that were most strongly associated with phenotypic variation within the approximately 1-Mb *Heliconius* interval. The diversity of wing patterning in *Heliconius* arises from divergence at wing pattern loci⁷, while convergent patterns generally involve the same loci and sometimes even the same alleles^{14–16}. We used this pattern of divergence and sharing to identify SNPs associated with colour

pattern elements across many individuals from a wide diversity of colour pattern phenotypes (Fig. 2).

In three separate *Heliconius* species, our analysis consistently implicated the gene *cortex* as being involved in adaptive differences in wing colour pattern. In *H. erato* the strongest associations with the presence of a yellow hindwing bar were centred around the genomic region

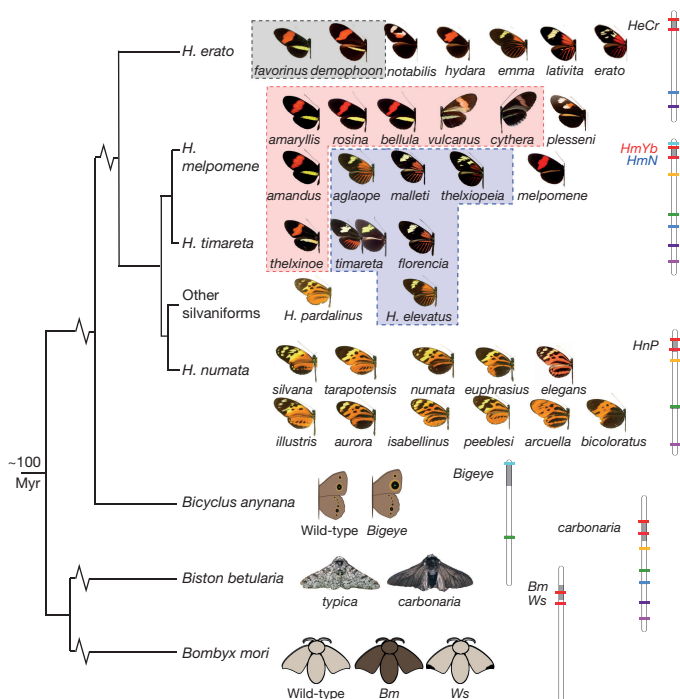


Figure 1 | A homologous genomic region controls a diversity of phenotypes across the Lepidoptera. Left, phylogenetic relationships²⁹. Right, chromosome maps with colour pattern intervals in grey; coloured bars represent markers used to assign homology^{5,8–10} and the first and last genes from Fig. 2 are shown in red. In *H. erato* the *HeCr* locus controls the yellow hindwing bar phenotype (grey boxed races). In *H. melpomene* it controls both the yellow hindwing bar (*HmYb*, pink box) and the yellow forewing band (*HmN*, blue box). In *H. numata* it modulates black, yellow and orange elements on both wings (*HnP*), producing phenotypes that mimic butterflies in the genus *Melinaea*. Morphs/races of *Heliconius* species included in this study are shown with names. All images are by the authors or are in the public domain.

¹Department of Animal and Plant Sciences, University of Sheffield, Western Bank, Sheffield, S10 2TN UK. ²Smithsonian Tropical Research Institute, Apartado Postal 0843-00153, Panamá, República de Panamá. ³Biology Program, Faculty of Natural Sciences and Mathematics, Universidad del Rosario, Cra. 24 No 63C-69, Bogotá D.C., 111221, Colombia. ⁴Institut de Systématique, Evolution et Biodiversité (UMR 7205 CNRS, MNHN, UPMC, EPHE, Sorbonne Université), Muséum National d'Histoire Naturelle, CP50, 57 rue Cuvier, 75005 Paris, France. ⁵Cell and Developmental Biology, John Innes Centre, Norwich, Norfolk NR4 7UH, UK. ⁶Research School of Biology, The Australian National University, 134 Linnaeus Way, Acton, ACT, 2601, Australia. ⁷Department of Zoology, University of Cambridge, Downing Street, Cambridge, CB2 3EJ, UK. ⁸Energy and Resources Group, University of California at Berkeley, California, 94720, USA. ⁹Department of Biology, Williams College, Williamstown, Massachusetts 01267, USA. ¹⁰Department of Zoology, University of Oxford, South Parks Rd, Oxford OX1 3PS, UK. ¹¹Penn State University, 517 Mueller, University Park, Pennsylvania 16802, USA. ¹²School of Biosciences, University of Exeter in Cornwall, Penryn, Cornwall TR10 9FE, UK. ¹³Centre d'Ecologie Fonctionnelle et Evolutive (CEFE, UMR 5175 CNRS, Université de Montpellier, Université Paul-Valéry Montpellier, EPHE), 1919 route de Mende, 34293 Montpellier, France.

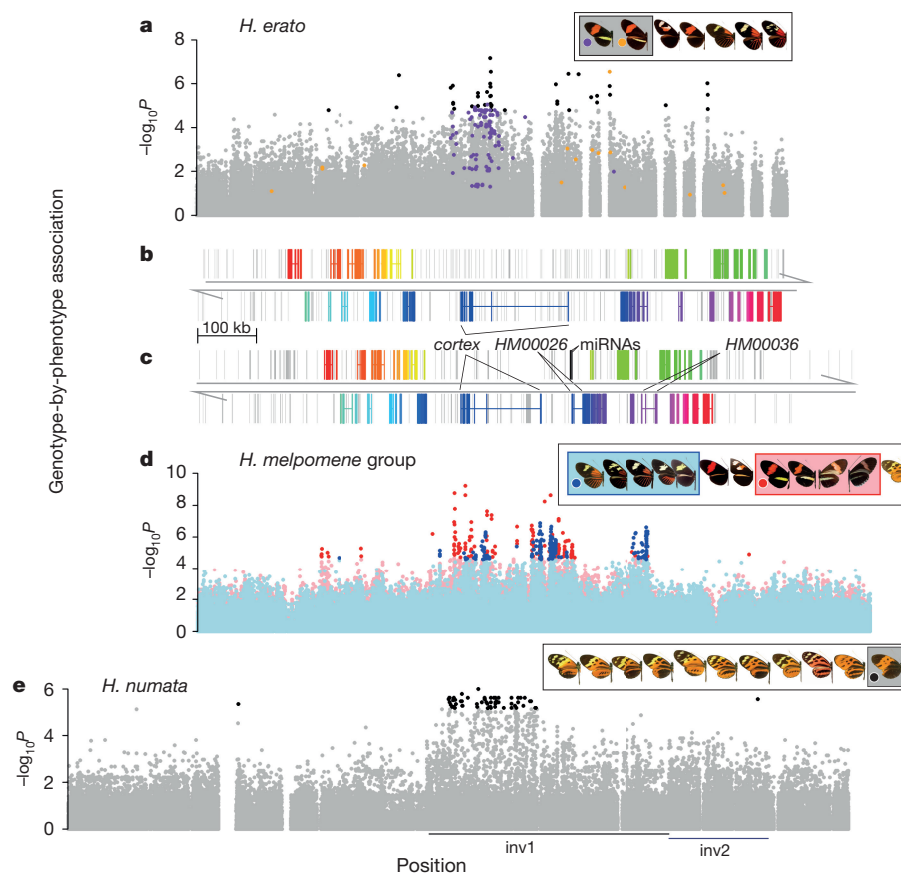


Figure 2 | Association analyses across the genomic region known to contain major colour pattern loci in *Heliconius*. **a**, Association in *H. erato* with the yellow hindwing bar ($n = 45$). Coloured SNPs are fixed for a unique state in *H. erato demophoon* (orange) or *H. erato favorinus* (purple). **b**, Genes in *H. erato* with direct homologues in *H. melpomene*. Genes are in different colours with exons (coding and UTRs) connected by lines. Grey bars are transposable elements. **c**, *H. melpomene* genes and transposable elements. Colours correspond to homologous *H. erato* genes and microRNAs³⁰ are black. **d**, Association in the *H. melpomene/timareta*/

silvaniform group with the yellow hindwing bar (red) and yellow forewing band (blue) ($n = 49$). **e**, Association in *H. numata* with the *bicoloratus* morph ($n = 26$); inversion positions¹³ shown below. In all cases black or dark coloured points are above the strongest associations found outside the colour pattern scaffolds (*H. erato* $P = 1.63 \times 10^{-5}$; *H. melpomene* $P = 2.03 \times 10^{-5}$ and $P = 2.58 \times 10^{-5}$ for hindwing bar and forewing band, respectively; *H. numata* $P = 6.81 \times 10^{-6}$). P values are from score tests for association.

containing *cortex* (Fig. 2a). We identified 108 SNPs that were fixed for one allele in *H. erato favorinus*, and fixed for the alternative allele in all individuals lacking the yellow bar; the majority of these SNPs were in introns of *cortex* (Extended Data Table 2). Fifteen SNPs showed a similar fixed pattern for *H. erato demophoon*, which also has a yellow bar. These SNPs did not overlap with those in *H. erato favorinus*, consistent with the hypothesis that this phenotype evolved independently in the two disjunct populations¹⁷.

Previous work has suggested that alleles at the *Yb* locus are shared between *H. melpomene*, the closely related species *H. timareta*, and the more distantly related species *H. elevatus*, resulting in mimicry among these species¹⁸. Across these species, the strongest associations with the yellow hindwing bar phenotype were again found at *cortex* (Fig. 2d, Extended Data Fig. 1a and Extended Data Table 3). Similarly, the strongest associations with the yellow forewing band were found around the 5' untranslated regions (UTRs) of *cortex* and *HM00036*, an orthologue of the *wash* gene in *Drosophila melanogaster*. A single SNP about 17 kb upstream of *cortex* (the closest gene) was perfectly associated with the yellow forewing band across all *H. melpomene*, *H. timareta* and *H. elevatus* individuals (Extended Data Figs 1a, 2 and Extended Data Table 3). We found no fixed coding sequence variants at *cortex* in larger samples (14–38 individuals) of *H. melpomene aglaope* and *H. melpomene amaryllis* (Extended Data Fig. 3 and Supplementary Information), which differ in *Yb*-controlled phenotypes¹⁹, suggesting that functional variants are likely to be regulatory rather

than coding. We found extensive transposable element variation around *cortex* but it is unclear whether any of these are associated with phenotypic differences (Extended Data Fig. 3, Extended Data Table 4 and Supplementary Information).

Finally, large inversions at the *P* supergene locus in *H. numata* (Fig. 1) are associated with different morphs¹³. There is a steep increase in genotype-by-phenotype association at the breakpoint of inversion 1, consistent with the role of these inversions in reducing recombination (Fig. 2e). However, the *bicoloratus* morph can recombine with all other morphs across one or the other inversion, permitting finer-scale association mapping of this region. As in *H. erato* and *H. melpomene*, this analysis showed a narrow region of associated SNPs corresponding exactly to the *cortex* gene (Fig. 2e), again with the majority of SNPs being found in introns (Extended Data Table 2). This associated region does not correspond to any other known genomic feature, such as an inversion or inversion breakpoint.

To determine whether sequence variants around *cortex* were regulating its expression, we investigated gene expression across the *Yb* locus. We used a custom designed microarray including probes from all predicted genes in the *H. melpomene* genome¹⁸ as well as probes tiled across the central portion of the *Yb* locus, focusing on two naturally hybridizing *H. melpomene* races (*plesseni* and *malleti*) that differ in *Yb*-controlled phenotypes⁷. *cortex* was the only gene across the entire interval to show significant differences in expression both between races with different wing patterns (false discovery rate (FDR)

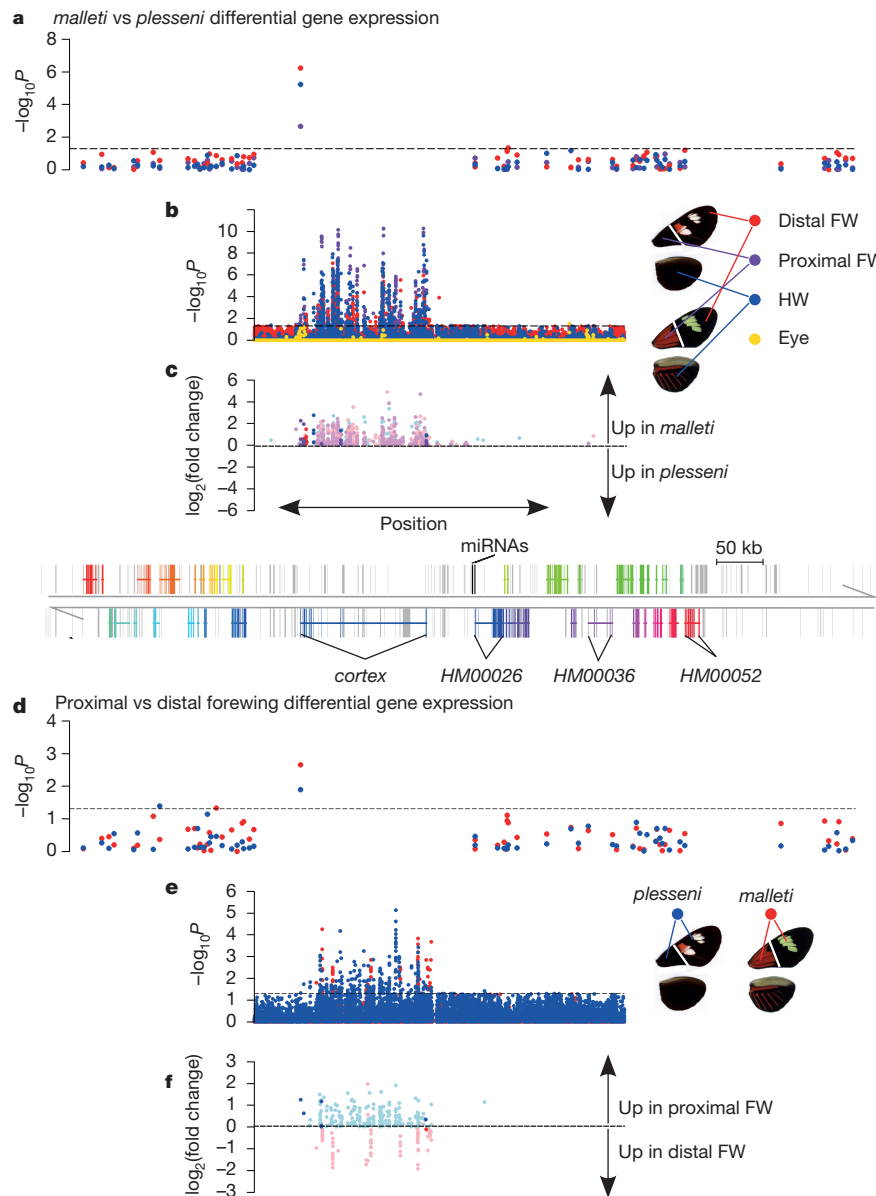


Figure 3 | Differential gene expression across the genomic region known to contain major colour pattern loci in *H. melpomene*.

a–f, Expression differences in day 3 pupae, for all genes in the *Yb* interval (**a**, **d**) and tiling probes spanning the central portion of the interval (**b**, **c**, **e**, **f**). Expression is compared between races for each wing region (**a–c**) and between proximal and distal forewing sections for each race

(**d–f**). **c**, **f**, Magnitude and direction of expression difference (\log_2 fold change) for tiling probes showing significant differences ($P \leq 0.05$); probes in known *cortex* exons shown in dark colours. Gene *HM00052* was differentially expressed between other races in RNA-seq data (Supplementary Information) but is not differentially expressed here. *P* values are based on FDR-adjusted *t*-statistics.

adjusted *t*-test $P = 6.09 \times 10^{-7}$) and between wing sections with different pattern elements (FDR adjusted *t*-test $P = 0.00224$; Fig. 3). This finding was reinforced in the tiled probe set, where we observed strong differences in the expression of *cortex* exons and introns but few differences outside this region (Extended Data Table 2). *cortex* expression was higher in *H. melpomene malleti* than in *H. melpomene plesseni* in all three wing sections used (but not eyes) (Fig. 3c and Extended Data Fig. 4c). When different wing sections were compared within each race, *cortex* expression in *H. melpomene malleti* was higher in the distal section that contains the *Yb*-controlled yellow forewing band than in the proximal section, consistent with *cortex* producing this band. In contrast, *H. melpomene plesseni*, which lacks the yellow band, had higher *cortex* expression in the proximal forewing section than in the distal section (Fig. 3f and Extended Data Fig. 4j). Differences in expression were found in pupal wings only on

days 1 and 3 but not on days 5 or 7 (Extended Data Fig. 4), similar to the pattern observed previously for the transcription factor *optix*²⁰.

Differential expression was not confined to the exons of *cortex*; the majority of differentially expressed probes in the tiling array corresponded to *cortex* introns (Fig. 3). This differential expression of introns does not appear to be due to transposable element variation (Extended Data Table 2), but may be due to elevated background transcription and unidentified splice variants. PCR with reverse transcription (RT-PCR) revealed a diversity of splice variants (Extended Data Fig. 5), and their sequenced products included eight non-constitutive exons and six variable donor/acceptor sites, but we did not exhaustively sequence all transcripts (Supplementary Information). We cannot rule out the possibility that some of the differentially expressed intronic regions could be distinct non-coding RNAs. However, quantitative RT-PCR (qRT-PCR) in other hybridizing races with

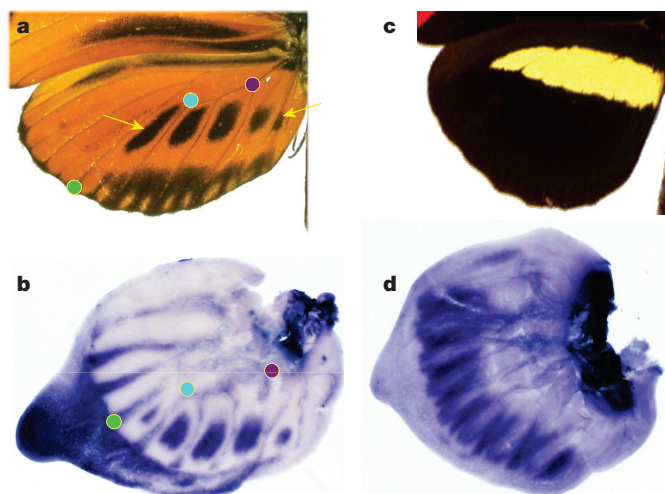


Figure 4 | *In situ* hybridizations of *cortex* in hindwings of final instar larvae. a, b, *H. numata tarapotensis* (replicated three times in the lab). Adult wing shown in a; coloured points indicate landmarks, yellow arrows highlight adult pattern elements corresponding to *cortex* staining. c, d, *H. melpomene rosina* (replicated twice in the lab). Adult wing shown in c; staining patterns in other *H. melpomene* races (*meriana*, $n = 11$, and *aglaope*, $n = 6$) appeared similar. The probe used was complementary to the *cortex* isoform with the longest open reading frame (also the most common; see Extended Data Fig. 5).

divergent *Yb* alleles (*aglaope/amaryllis* and *rosina/melpomene*) also identified differences in *cortex* expression and allele-specific splicing differences between both pairs of races (Extended Data Figs 1, 5 and Supplementary Information).

Finally, *in situ* hybridization of *cortex* in final instar larval hindwing discs showed expression in wing regions fated to become black in the adult wing, most strikingly in their correspondence to the black patterns on adult *H. numata* wings (Fig. 4). In contrast, the array results from pupal wings were suggestive of higher expression in non-melanic regions. This may suggest that *cortex* is upregulated at different time-points in wing regions fated to become different colours.

Overall, *cortex* shows significant differential expression and is the only gene in the candidate region to be consistently differentially expressed in multiple race comparisons and between differently patterned wing regions. Coupled with the strong genotype-by-phenotype associations across multiple independent lineages (Extended Data Table 1), these findings strongly implicate *cortex* as a major regulator of colour and pattern. However, we have not excluded the possibility that other genes in this region also influence pigmentation patterning. A prominent role for *cortex* is also supported by studies in other taxa; our identification of distant 5' untranslated exons of *cortex* (Supplementary Information) suggests that the 100-bp interval containing the *Ws* mutation in *B. mori* is likely to be within an intron of *cortex* and not in intergenic space, as previously thought¹⁰. In addition, fine mapping and gene expression also suggest that *cortex* controls melanism in the peppered moth⁴.

It seems likely that *cortex* controls pigmentation patterning by controlling scale cell development. The *cortex* gene falls in an insect-specific lineage within the *fzy* (also known as *Cdc20/fizzy*) family of cell-cycle regulators (Extended Data Fig. 6a). The phylogenetic tree of this gene family highlighted three major orthologous groups, two of which have highly conserved functions in cell-cycle regulation, mediated through interaction with the anaphase-promoting complex/cyclosome (APC/C)^{3,21}. The third group, containing *cortex* proteins, is evolving rapidly, with low amino acid identity between *D. melanogaster* and *H. melpomene cortex* (14.1%), contrasting with much higher identities for orthologues between these species in the other two groups (*fzy*, 47.8% and *rap* (also known as *fzr*, *cdh1*, *rap/Fzr*), 47.2%; Extended Data Fig. 6a). *D. melanogaster cortex* acts through a similar

mechanism to *fzy* to control meiosis in the female germ line^{22–24}. *H. melpomene cortex* also has some conservation of the fizzy family C-box and IR (isoleucine–arginine) tail elements (Supplementary Information) that mediate binding to the APC/C²³, suggesting that it may have retained a cell-cycle function, although we found that expressing *H. melpomene cortex* in *D. melanogaster* wings produced no detectable effect (Extended Data Fig. 6 and Supplementary Information).

Previously identified butterfly wing patterning genes have been transcription factors or signalling molecules^{20,25}. Developmental rate has long been thought to play a role in lepidopteran patterning^{26,27}, but *cortex* was not a likely a priori candidate, because its *Drosophila* orthologue has a highly specific function in meiosis²³. The recruitment of *cortex* to wing patterning appears to have occurred before the major diversification of the Lepidoptera and this gene has repeatedly been targeted by natural selection^{1,7,9,28} to generate both cryptic⁴ and aposematic patterns.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 3 July 2015; accepted 29 March 2016.

1. Cook, L. M., Grant, B. S., Saccheri, I. J. & Mallet, J. Selective bird predation on the peppered moth: the last experiment of Michael Majerus. *Biol. Lett.* **8**, 609–612 (2012).
2. Jiggins, C. D. Ecological speciation in mimetic butterflies. *Bioscience* **58**, 541–548 (2008).
3. Dawson, I. A., Roth, S. & Artavanis-Tsakonas, S. The *Drosophila* cell cycle gene *fizzy* is required for normal degradation of cyclins A and B during mitosis and has homology to the *CDC20* gene of *Saccharomyces cerevisiae*. *J. Cell Biol.* **129**, 725–737 (1995).
4. van't Hof, A. E. et al. The industrial melanism mutation in British peppered moths is a transposable element. *Nature* <http://dx.doi.org/10.1038/nature17951> (this issue).
5. Joron, M. et al. A conserved supergene locus controls colour pattern diversity in *Heliconius* butterflies. *PLoS Biol.* **4**, e303 (2006).
6. Sheppard, P. M., Turner, J. R. G., Brown, K. S., Benson, W. W. & Singer, M. C. Genetics and the evolution of Müllerian mimicry in *Heliconius* butterflies. *Phil. Trans. R. Soc. Lond. B* **308**, 433–610 (1985).
7. Nadeau, N. J. et al. Population genomics of parallel hybrid zones in the mimetic butterflies, *H. melpomene* and *H. erato*. *Genome Res.* **24**, 1316–1333 (2014).
8. Beldade, P., Saenko, S. V., Pul, N. & Long, A. D. A. Gene-based linkage map for *Bicyclus anynana* butterflies allows for a comprehensive analysis of synteny with the lepidopteran reference genome. *PLoS Genet.* **5**, e1000366 (2009).
9. van't Hof, A. E., Edmonds, N., Dalíková, M., Marec, F. & Saccheri, I. J. Industrial melanism in British peppered moths has a singular and recent mutational origin. *Science* **332**, 958–960 (2011).
10. Ito, K. et al. Mapping and recombination analysis of two moth colour mutations, Black moth and Wild wing spot, in the silkworm *Bombyx mori*. *Heredity* **116**, 52–59 (2016).
11. Counterman, B. A. et al. Genomic hotspots for adaptation: the population genetics of Müllerian mimicry in *Heliconius erato*. *PLoS Genet.* **6**, e1000796 (2010).
12. Ferguson, L. et al. Characterization of a hotspot for mimicry: assembly of a butterfly wing transcriptome to genomic sequence at the *HmYb/Sb* locus. *Mol. Ecol.* **19**, 240–254 (2010).
13. Joron, M. et al. Chromosomal rearrangements maintain a polymorphic supergene controlling butterfly mimicry. *Nature* **477**, 203–206 (2011).
14. Hines, H. M. et al. Wing patterning gene redefines the mimetic history of *Heliconius* butterflies. *Proc. Natl Acad. Sci. USA* **108**, 19666–19671 (2011).
15. Pardo-Díaz, C. et al. Adaptive introgression across species boundaries in *Heliconius* butterflies. *PLoS Genet.* **8**, e1002752 (2012).
16. Wallbank, R. W. R. et al. Evolutionary novelty in a butterfly wing pattern through enhancer shuffling. *PLoS Biol.* **14**, e1002353 (2016).
17. Maroja, L. S., Alschuler, R., McMillan, W. O. & Jiggins, C. D. Partial complementarity of the mimetic yellow bar phenotype in *Heliconius* butterflies. *PLoS ONE* **7**, e48627 (2012).
18. The *Heliconius* Genome Consortium. Butterfly genome reveals promiscuous exchange of mimicry adaptations among species. *Nature* **487**, 94–98 (2012).
19. Mallet, J. The genetics of warning colour in peruvian hybrid zones of *Heliconius erato* and *H. melpomene*. *Proc. R. Soc. Lond. B* **236**, 163–185 (1989).
20. Reed, R. D. et al. Optix drives the repeated convergent evolution of butterfly wing pattern mimicry. *Science* **333**, 1137–1141 (2011).
21. Barford, D. Structural insights into anaphase-promoting complex function and mechanism. *Philos. Trans. R. Soc. B* **366**, 3605–3624 (2011).

22. Chu, T., Henrion, G., Haegeli, V. & Strickland, S. *Cortex*, a *Drosophila* gene required to complete oocyte meiosis, is a member of the Cdc20/fizzy protein family. *Genesis* **29**, 141–152 (2001).
23. Pesin, J. A. & Orr-Weaver, T. L. Developmental role and regulation of cortex, a meiosis-specific anaphase-promoting complex/cyclosome activator. *PLoS Genet.* **3**, e202 (2007).
24. Swan, A. & Schüpbach, T. The Cdc20/Cdh1-related protein, Cort, cooperates with Cdc20/Fzy in cyclin destruction and anaphase progression in meiosis I and II in *Drosophila*. *Development* **134**, 891–899 (2007).
25. Martin, A. *et al.* Diversification of complex butterfly wing patterns by repeated regulatory evolution of a Wnt ligand. *Proc. Natl Acad. Sci. USA* **109**, 12632–12637 (2012).
26. Koch, P. B., Lorenz, U., Brakefield, P. M. & French-Constant, R. H. Butterfly wing pattern mutants: developmental heterochrony and co-ordinately regulated phenotypes. *Dev. Genes Evol.* **210**, 536–544 (2000).
27. Gilbert, L. E., Forrest, H. S., Schultz, T. D. & Harvey, D. J. Correlations of ultrastructure and pigmentation suggest how genes control development of wing scales of *Heliconius* butterflies. *J. Res. Lepid.* **26**, 141–160 (1988).
28. Mallet, J. & Barton, N. H. Strong natural selection in a warning-color hybrid zone. *Evolution* **43**, 421–431 (1989).
29. Wahlberg, N., Wheat, C. W. & Peña, C. Timing and patterns in the taxonomic diversification of Lepidoptera (butterflies and moths). *PLoS ONE* **8**, e80875 (2013).
30. Surridge, A. K. *et al.* Characterisation and expression of microRNAs in developing wings of the neotropical butterfly *Heliconius melpomene*. *BMC Genomics* **12**, 62 (2011).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank C. Sasaki for assembly of the *H. erato* BACs; M. Abanto and A. Tapia for assistance with raising butterflies; M. Chouteau, J. Morris and K. Dasmahapatra for providing larvae for *in situ* hybridizations; A. Morrison, R. Tetley, S. Carl and H. Wegener for assistance with laboratory work; S. Baxter for the *H. melpomene* fosmid libraries; and the governments

of Colombia, Ecuador, Panama and Peru for permission to collect butterflies. This work was funded by a Leverhulme Trust award (RPG-2014-167), BBSRC (H01439X/1), ERC (SpeciationGenetics 339873), and NERC small project (MGF 280) grants to C.D.J., NSF grants (DEB 1257689, IOS 1052541) to W.O.M., an ERC starting grant (StG-243179) to M.J. and French National Agency for Research (ANR) grants to M.J. (ANR-12-JSV7-0005) and V.L. (ANR-13-JSV7-0003-01). N.J.N. is funded by a NERC fellowship (NE/K008498/1).

Author Contributions N.J.N. performed the association analyses, 5' RACE, RT-PCR and qRT-PCR and prepared the manuscript. N.J.N. and C.D.J. co-ordinated the research. C.P.-D. performed and analysed the microarray and RNA-seq experiments. A.W. performed the *H. numata* association analysis. M.A.S. assembled and annotated the *HeCr* BAC reference and the *H. erato* alignments. S.V.S. performed *in situ* hybridizations. R.W.R.W. performed the transgenic experiments and analysis of *de novo* assembled sequences and fosmids together with J.J.H. G.C.W. and L.F. initially identified splicing variants of *cortex*. L.M. performed crosses between *H. melpomene* races. H.H. screened the *HeCr* BAC library. C.S. and R.M.M. provided samples. A.J.D. contributed to the *H. melpomene* BAC sequencing and annotation. R.H.f.-C., M.J., V.L., W.O.M. and C.D.J. are PIs who obtained funding and led the project elements. All authors commented on the manuscript.

Author Information Short read sequence data generated for this study are available from ENA (<http://www.ebi.ac.uk/ena>) under study accession PRJEB8011 and PRJEB12740 (see Supplementary Table 1 for previously published data accessions). The updated *Cr* contig is deposited in Genbank with accession KC469893.2. The assembled *H. melpomene* fosmid sequences are deposited in Genbank with accessions KU514430–KU514438. The microarray data are deposited in GEO with accessions GSM1563402–GSM1563497. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to N.J.N. (n.nadeau@sheffield.ac.uk) or C.D.J. (c.jiggins@zoo.cam.ac.uk).

METHODS

No statistical methods were used to predetermine sample size.

***H. erato* Cr reference.** *Cr* is the homologue of *Yb* in *H. erato* (Fig. 1). An existing reference for this region was available in three pieces³¹ (467,734 bp, 114,741 bp and 161,149 bp; GenBank KC469893.1). We screened the same bacterial artificial chromosome (BAC) library used previously^{11,31} using described procedures¹¹ with probes designed to the ends of the existing BAC sequences and the *HmYb* BAC reference sequence. Two BACs (04B01 and 10B14) were identified as spanning one of the gaps and sequenced using Illumina 2 × 250-bp paired-end reads collected on the Illumina MiSeq. The raw reads were screened to remove vector and *Escherichia coli* bases. The first 50,000 read pairs were taken for each BAC and assembled individually with the Phrap³² software and manually edited with consed³³. Contigs with discordant read pairs were manually broken and properly merged using concordant read data. Gaps between contig ends were filled using an in-house finishing technique in which the terminal 200 bp of the contig ends were extracted and queried against the unused read data for spanning pairs, which were added using the addSolexaReads.perl script in the consed package. Finally, a single reference contig was generated by identifying and merging overlapping regions of the two consensus BAC sequences.

To fill the remaining gap (between positions 800,387 and 848,446) we used the overhanging ends to search the scaffolds from a preliminary *H. erato* genome assembly of five Illumina paired-end libraries with different insert sizes (250, 500, 800, 4,300 and 6,500 bp) from two related *H. erato* demophoon individuals. We identified two scaffolds (scf1869 and scf1510) that overlapped and spanned the gap (using 12,257 bp of the first scaffold and 35,803 bp of the second).

The final contig was 1,009,595 bp in length, of which 2,281 bp were unknown (N). The *HeCr* assembly was verified by aligning to the *HmYb* genome scaffold (HE667780) with mummer and blast. The *HeCr* contig was annotated as described previously³¹ with some minor modifications. Briefly this involved first generating a reference-based transcriptome assembly with existing *H. erato* RNA-sequenced (RNA-seq) wing tissue (GenBank accession SRA060220). We used Trimmomatic³⁴ (v0.22), and FLASH³⁵ (v1.2.2) to prepare the raw sequencing reads, checking the quality with FastQC³⁶ (v0.10.0). We then used the Bowtie/TopHat/Cufflinks^{37–39} pipeline to generate transcripts for the unmasked reference sequence. We generated gene predictions with the MAKER pipeline⁴⁰ (v2.31). Homology and synteny in gene content with the *H. melpomene* *Yb* reference were identified by aligning the *H. melpomene* coding sequences to the *H. erato* reference with BLAST. Homologous genes were present in the same order and orientation in *H. erato* and *H. melpomene* (Fig. 2b, c). Annotations were manually adjusted if genes had clearly been merged or split in comparison to *H. melpomene* (which has been extensively manually curated¹²). In addition, *H. erato* *cortex* was manually curated from the RNA-seq data and using Exonerate⁴¹ alignments of the *H. melpomene* protein and mRNA transcripts, including the 5' UTRs.

Genotype-by-phenotype association analyses. Information on the individuals used and ENA accessions for sequence data are given in Supplementary Table 1. We used shotgun Illumina sequence reads from 45 *H. erato* individuals from 7 races that were generated as part of a previous study³¹ (Supplementary Information). Reads were aligned to an *H. erato* reference containing the *Cr* contig and other sequenced *H. erato* BACs^{11,31} with BWA⁴², which has previously been found to work better than Stampy⁴³ (which was used for the alignments in the other species) with an incomplete reference sequence³¹. The parameters used were as follows: maximum edit distance (n), 8; maximum number of gap opens (o), 2; maximum number of gap extensions (e), 3; seed (l), 35; maximum edit distance in seed (k), 2. We then used Picard tools to remove PCR and optical duplicate sequence reads and GATK⁴⁴ to re-align indels and call SNPs using all individuals as a single population. Expected heterozygosity was set to 0.2 in GATK. 132,397 SNPs were present across *Cr*. A further 52,698 SNPs not linked to colour pattern loci were used to establish background association levels.

For the *H. melpomene*/*H. numata* clade we used previously published sequence data from 19 individuals from enrichment sequencing targeting the *Yb* region, the unlinked *HmB/D* region that controls the presence or absence of red colour pattern elements, and ~1.8 Mb of non-colour pattern genomic regions⁴⁵, as well as 9 whole-genome shotgun-sequenced individuals^{18,46}. We added targeted sequencing and shotgun whole-genome sequencing of an additional 47 individuals (Supplementary Information). Alignments were performed using Stampy⁴³ with default parameters except for substitution rate which was set to 0.01. We again removed duplicates and used GATK to re-align indels and call SNPs with expected heterozygosity set to 0.1.

The analysis of *H. melpomene*/*timareta*/*silvaniform* included 49 individuals, which were aligned to v1.1 of the *H. melpomene* reference genome with the scaffolds containing *Yb* and *HmB/D* swapped with reference BAC sequences¹⁸, which contained fewer gaps of unknown sequence than the genome scaffolds. The *Yb*

region contained 232,631 SNPs and a further 370,079 SNPs were used to establish background association levels.

The *H. numata* analysis included 26 individuals aligned to unaltered v1.1 of the *H. melpomene* reference genome, because the genome scaffold containing *Yb* is longer than the BAC reference, making it easier to compare the inverted and non-inverted regions in this species. We tested for associations at 262,137 SNPs on the *Yb* scaffold with the *H. numata* *bicoloratus* morph, which had a sample size of 5 individuals.

We measured associations between genotype and phenotype using a score test (qtscore) in the GenABEL package in R (ref. 47). This was corrected for background population structure using a test specific inflation factor (λ) calculated from the SNPs unlinked to the major colour pattern controlling loci (described above), as the colour pattern loci are known to have a different population structure from the rest of the genome^{14,15,18}. We used a custom perl script to convert GATK vcf files to Illumina SNP format for input to GenABEL⁴⁷. GenABEL does not accept multiallelic sites, so the script also converted the genotype of any individuals for which a third (or fourth) allele was present to a missing genotype (with these defined as the lowest frequency alleles). Custom R scripts were used to identify sites showing perfect associations with calls for >75% of individuals.

Microarray gene expression analyses. We designed a Roche NimbleGen microarray (12 × 135K format) with probes for all annotated *H. melpomene* genes¹⁸ and tiling of the central portion of the *Yb* BAC sequence contig that was previously identified as showing the strongest differentiation between *H. melpomene* races⁴⁵. In addition to the *HmYb* tiling array probes there were 6,560 probes tiling *HmA*c (a third unlinked colour pattern locus) and 10,716 probes tiling *HmB/D*, again distanced on average at 10-bp intervals. The whole-genome gene expression array contained 107,898 probes in total.

This array was interrogated with Cy3-labelled double-stranded cDNA generated from total RNA (with a SuperScript double-stranded cDNA synthesis kit (Invitrogen) and a one-colour DNA labelling kit (Nimagen)) from four pupal developmental stages of *H. melpomene plesseni* and *malleti*. Pupae were from captive stocks maintained in insectary facilities in Gamboa, Panama. Tissue was stored in RNA later (Ambion) at –80°C before RNA extraction. RNA was extracted using TRIzol (Invitrogen) followed by purification with RNeasy (Qiagen) and DNase treated with DNA-free (Ambion). Quantification was performed using a Qubit 2.0 fluorometer (Invitrogen) and purity and integrity assessed using a Bioanalyzer 2100 (Agilent). Samples were randomized and each hybridized to a separate array. The *HmYb* probe array contained 9,979 probes distanced on average at 10 bp. The whole-genome expression array contained on average 9 probes per annotated gene in the genome (v1.1 (ref. 18)) as well as any transcripts not annotated but predicted from RNA-seq evidence.

Background corrected expression values for each probe were extracted using NimbleScan software (v2.3). Analyses were performed with the LIMMA package implemented in R/Bioconductor⁴⁸. The tiling array and whole-genome data sets were analysed separately. Expression values were extracted and quantile-normalized, log₂-transformed, quality controlled and analysed for differences in expression between individuals and wing regions. *P* values were adjusted for multiple hypothesis testing using the false discovery rate (FDR) method⁴⁹.

In situ hybridization. *H. numata* and *H. melpomene* larvae were reared in a greenhouse at 25–30°C and sampled at the last instar. *In situ* hybridizations were performed according to previously described methods²⁵ with a *cortex* riboprobe synthesized from a 831-bp cDNA amplicon from *H. numata*. Wing discs were incubated in a standard hybridization buffer containing the probe for 20–24 h at 60°C. For secondary detection of the probe, wing discs were incubated in a 1:3,000 dilution of anti-digoxigenin alkaline phosphatase Fab fragments and stained with BM Purple for 3–6 h at room temperature. Stained wing discs were photographed with a Leica DFC420 digital camera mounted on a Leica Z6 APO stereomicroscope.

De novo assembly of short read data in *H. melpomene* and related taxa. To better characterize indel variation from the short-read sequence data used for the genotype-by-phenotype association analysis, we performed *de novo* assemblies of a subset of *H. melpomene* individuals and related taxa with a diversity of phenotypes (Extended Data Fig. 2). Assemblies were performed using the *de novo* assembly function of CLCGenomics Workbench v6.0 under default parameters. The assembled contigs were then BLASTed against the *Yb* region of the *H. melpomene melpomene* genome¹⁸, using Geneious v8.0. The contigs identified by BLAST were then concatenated to generate an allele sequence for each individual. Occasionally two unphased alleles were generated when two contigs were matched to a given region. If more than two contigs of equal length matched then this was considered an unresolvable repeat region and replaced with Ns. The assembled alleles were then aligned using the MAFFT alignment plugin in Geneious v8.0.

Long-range PCR targeted sequencing of *cortex* in *H. melpomene aglaope* and *H. melpomene amaryllis*. We generated two long-range PCR products covering

88.8% of the 1,344-bp coding region of *cortex* (excluding 67 bp at the 5' end and 83 bp at the 3' end; see Supplementary Information). A product spanning coding exons 5–9 (the final exon) was obtained from 29 *H. melpomene amaryllis* individuals and 29 *H. melpomene aglaope* individuals; a product spanning coding exons 2–5 was obtained from 32 *H. melpomene amaryllis* individuals and 14 *H. melpomene aglaope* individuals. In addition, a product spanning exons 4–6 was obtained from six *H. melpomene amaryllis* and five *H. melpomene aglaope* individuals that failed to amplify one or both of the larger products. Long-range PCR was performed using Extensor long-range PCR mastermix (Thermo Scientific) following the manufacturer's guidelines with a 60 °C annealing temperature in a 10–20- μ l volume. The product spanning coding exons 5–9 was obtained with primers HM25_long_F1 and HM25_long_R4 (see Supplementary Table 2 for primer sequences); the product spanning coding exons 2–5 was obtained with primers HM25_long_F4 and HM25_long_R2; the product spanning exons 4–6 was obtained with primers 25_ex5-ex7_r1 and 25_ex5-ex7_f1. Products were pooled for each individual, including five additional products from the *Yb* locus and seven products in the region of the *Hmb/D* locus. They were then cleaned using QIAquick PCR purification kit (QIAGEN) before being quantified with a Qubit Fluorometer (Life Technologies) and pooled in equimolar amounts for each individual, taking into account variation in the length and number of PCR products included for each individual (because of some PCR failures, that is, proportionally less DNA was included if some PCR products were absent for a given individual).

Products were pooled within individuals (including additional products for other genes not analysed here) and then quantified and pooled in equimolar amounts for each individual within each race. The pooled products for each race (*H. melpomene aglaope* and *amaryllis*) were then prepared as two separate libraries with molecular identifiers and sequenced on a single lane of an Illumina GAIIx. Analysis was performed using Galaxy and the history is available at <https://usegalaxy.org/u/njnadeau/h/long-pcr-final>. Reads were quality filtered with a minimum quality of 20 required over 90% of the read, which resulted in 5% of reads being discarded. Reads were then quality trimmed to remove bases with quality less than 20 from the ends. They were then aligned to the target regions using the fosmid sequences from known races⁴⁵ with sequence from the *Yb* BAC walk¹² used to fill any gaps. Alignments were performed with BWA v.0.5.6 (ref. 42) and converted to pileup format using Samtools v.0.1.12 before being filtered on the basis of quality (≥ 20) and coverage (≥ 10). BWA alignment parameters were as follows: fraction of missing alignments given 2% uniform base error rate (aln -n) 0.01; maximum number of gap opens (aln -o) 2; maximum number of gap extensions (aln -e) 12; disallow long deletion within 12 bp towards the 3'-end (aln -d); number of first subsequences to take as seed (aln -l) 100. We then calculated coverage and minor allele frequencies for each race and the difference between these using custom scripts in R⁵⁰.

Sequencing and analysis of *H. melpomene* fosmid clones. Fosmid libraries had previously been made from single individuals of three *H. melpomene* races (*rosina*, *amaryllis* and *aglaope*) and several clones overlapping the *Yb* interval had been sequenced⁴⁵. We extended the sequencing of this region, particularly the region overlapping *cortex*, by sequencing an additional four clones from *H. melpomene rosina* (1051_83D21, accession KU514430; 1051_97A3, accession KU514431; 1051_65N6, accession KU514432; 1051_93D23, accession KU514433), two clones from *H. melpomene amaryllis* (1051_13K4, accession KU514434; 1049_8P23, accession KU514435) and three clones from *H. melpomene aglaope* (1048_80B22, accession KU514437; 1049_19P15, accession KU514436; 1048_96A7, accession KU514438). These were sequenced on a MiSeq 2000, and assembled using the *de novo* assembly function of CLCGenomics Workbench v.6.0. The individual clones (including existing clones 1051-143B3, accession FP578990; 1049-27G11, accession FP700055; and 1048-62H20, accession FP565804) were then aligned to the BAC and genome scaffold¹⁸ references using the MAFFT alignment plugin of Geneious v.8.0. Regions of general sequence similarity were identified and visualized using MAUVE⁵¹. We merged overlapping clones from the same individual if they showed no sequence differences, indicating that they came from the same allele. We identified transposable elements using nBLAST with an insect transposable element list downloaded from Repbase Update⁵², including known *Heliconius*-specific transposable elements⁵³.

5' RACE, RT-PCR and qRT-PCR. All tissues used for gene expression analyses were dissected from individuals from captive stocks derived from wild-caught individuals of various races of *H. melpomene* (*aglaope*, *amaryllis*, *melpomene*, *rosina*, *plesseni* and *malleti*) and F₂ individuals from a *H. melpomene rosina* (female) \times *H. melpomene melpomene* (male) cross. Experimental individuals were reared at 28–31 °C. Developing wings were dissected and stored in RNAlater (Ambion Life Technologies). RNA was extracted using a QIAgen RNeasy Mini kit following the manufacturer's guidelines and treated with TURBO DNA-free DNase kit (Ambion Life Technologies) to remove remaining genomic DNA. RNA

quantification was performed with a Nanodrop spectrophotometer, and the RNA integrity was assessed using the Bioanalyzer 2100 system (Agilent).

Total RNA was thoroughly checked for DNA contamination by performing PCR for EF1 α (using primers efl-a_RT_for and efl-a_RT_rev, Supplementary Table 2) with 0.5 μ l of RNA extract (50 ng–1 μ g of RNA) in a 20- μ l reaction using a polymerase enzyme that is not functional with RNA template (BioScript, Boline Reagents Ltd). If a product amplified within 45 cycles then the RNA sample was re-treated with DNase.

Single-stranded cDNA was synthesized using BioScript MMLV Reverse Transcriptase (Boline Reagents Ltd) with random hexamer (N6) primers and 1 μ g of template RNA from each sample in a 20- μ l reaction volume following the manufacturer's protocol. The resulting cDNA samples were then diluted 1:1 with nuclease-free water and stored at –80 °C.

5' RACE (rapid amplification of cDNA ends) was performed using RNA from hindwing discs from one *H. melpomene aglaope* and one *H. melpomene amaryllis* final instar larvae with a SMARTer RACE kit from Clontech. The gene-specific primer used for the first round of amplification was anchored in exon 4 (fzl_raceex5_R1; Supplementary Table 2). Secondary PCR of these products was then performed using a primer in exon 2 (HM25_long_F2; Supplementary Table 2) and the nested universal primer A. Other isoforms were detected by RT-PCR using primers within exons 2 and 9 (gene25_for_full1 and gene25_rev_ex3). We identified isoforms from 5' RACE and RT-PCR products by cutting individual bands from agarose gels and if necessary by cloning products before Sanger sequencing. Cloning of products was performed using TOPO TA (Invitrogen) or pGEM-T (Promega) cloning kits. Sanger sequencing was performed using BigDye terminator v3.1 (Applied Biosystems) run on an ABI13730 capillary sequencer. Primers fzl_ex1a_F1 and fzl_ex4_R1 were used to confirm expression of the furthest 5' UTR. For isoforms that appeared to show some degree of race specificity, we designed isoform-specific PCR primers spanning specific exon junctions (Extended Data Figs 2, 4 and Supplementary Table 2) and used these to either qualitatively (RT-PCR) or quantitatively (qRT-PCR) assess differences in expression between races.

We performed qRT-PCR using SensiMix SYBR green (Boline Reagents Ltd) with 0.2–0.25 μ M of each primer and 1 μ l of the diluted product from the cDNA reactions. Reactions were performed in an Opticon 2 DNA engine (MJ Research) with the following cycling parameters: 95 °C for 10 min; 35–50 \times (95 °C for 15 s, 55–60 °C for 30 s, 72 °C for 30 s); 72 °C for 5 min. Melting curves were generated between 55 °C and 90 °C with readings taken every 0.2 °C for each of the products to check that a single product was generated. At least one product from each set of primers was also run on a 1% agarose gel to check that a single product of the expected size was produced and the identity of the product was confirmed by direct sequencing (see Supplementary Table 2 for details of primers for each gene). We used two housekeeping genes (*EF1A* and *RPS3A*) for normalization and all results were taken as averages of triplicate PCR reactions for each sample.

C_t values were defined as the point at which fluorescence crossed a threshold (R_{ct}) adjusted manually to be the point at which fluorescence rose above the background level. Amplification efficiencies (E) were calculated using a dilution series of clean PCR products. Starting fluorescence, which is proportional to the starting template quantity, was calculated as $R_0 = R_{ct}(1 + E)^{-C_t}$. Normalized values were then obtained by dividing R_0 values for the target loci by R_0 values for *EF1A* and *RPS3A*. Results from both of these controls were always very similar, so the results presented are normalized to the mean of *EF1A* and *RPS3A*. All results were taken as averages of triplicate PCR reactions. If one of the triplicate values was more than one cycle away from the mean then this replicate was excluded. Similarly any individuals that were more than two s.d. away from the mean of all individuals for the target or normalization genes were excluded (these are not included in the numbers of individuals reported). Statistical significance was assessed by Wilcoxon rank sum tests performed in R (ref. 50).

RNA-seq analysis of *H. melpomene amaryllis/aglaope*. RNA-seq data for hindwings from three developmental stages had previously been obtained for two individuals of each race at each stage (12 individuals in total) and used in the annotation of the *H. melpomene* genome¹⁸ (deposited in ENA under study accessions ERP000993 and PRJEB7951). Four samples were multiplexed on each sequencing lane with the fifth instar larval and day 2 pupal samples sequenced on a GAIIx sequencer and the day 3 pupal wings sequenced on a HiSeq 2000 sequencer.

Two methods were used for alignment of reads to the reference genome and inferring read counts: Stampy⁴³ and RSEM (RNaseq by Expectation Maximisation)⁵⁴. In addition we used two different R/Bioconductor packages for estimation of differential gene expression: DESeq⁵⁵ and BaySeq⁵⁶. Read bases with quality scores <20 were trimmed with FASTX-Toolkit (http://hannonlab.cshl.edu/fastx_toolkit/index.html). Stampy was run with default parameters except for mean insert size, which was set to 500, s.d. 100, and substitution rate, which was set to

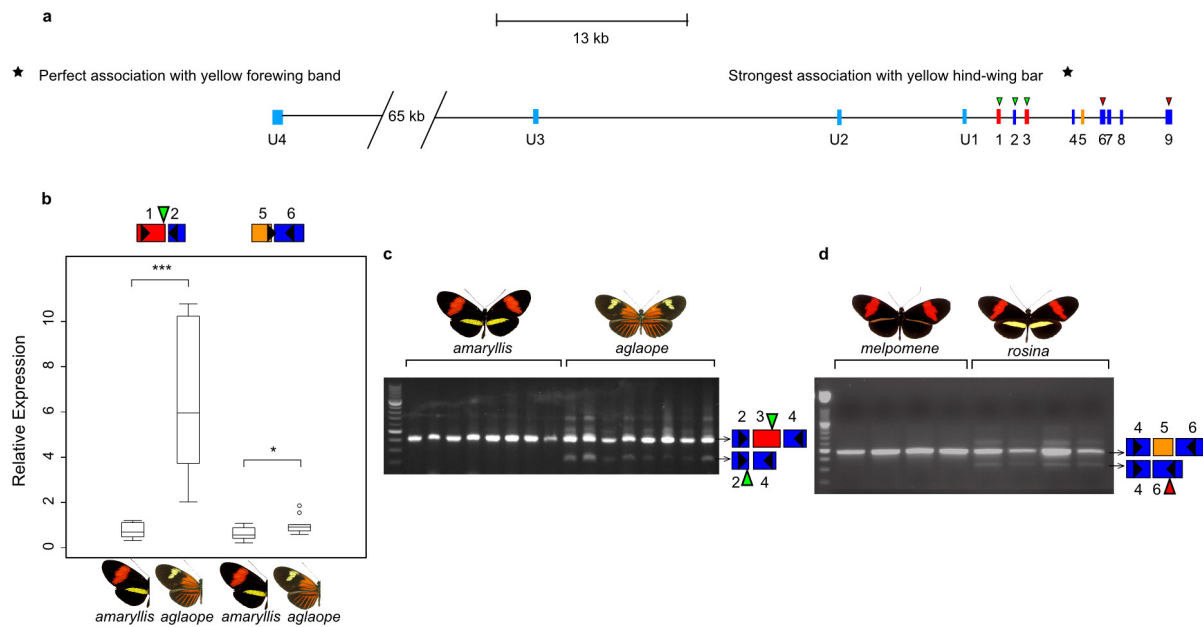
0.01. Alignments were filtered to exclude reads with mapping quality <30 and sorted using Samtools⁵⁷. We used the HT seq-count script with HTseq⁵⁸ to infer counts per gene from the BAM files.

RSEM⁵⁴ was run with default parameters to infer a transcriptome and then map RNA-seq reads against this using Bowtie³⁷ as an aligner. This was run with default parameters except for the maximum number of mismatches, which was set to 3.

Annotation and alignment of fizzy family proteins. In the arthropod genomes, some fizzy family proteins were found to be poorly annotated based on alignments to other family members. In these cases annotations were improved using well-annotated proteins from other species as references in the program Exonerate⁴¹ and the outputs were manually curated. Specifically, the annotation of *B. mori rap* (also known as *fzr*) was extended based on alignment of *Danaus plexippus rap*; the annotation of *B. mori fzy* was altered based on alignment of *D. melanogaster* and *D. plexippus fzy*; *H. melpomene fzy* was identified as part of the annotated gene HMEL017486 on scaffold HE671623 (Hmel v.1.1) based on alignment of *D. plexippus fzy*; the *Apis mellifera rap* annotation was altered based on alignment of *D. melanogaster rap*; the annotation of *Acyrtosiphon pisum rap* was altered based on alignment of *D. melanogaster rap*. No one-to-one orthologues of *D. melanogaster fzr2* were found in any of the other arthropod genera, suggesting that this gene is *Drosophila*-specific. Multiple sequence alignment of all the fizzy family proteins was then performed using the Expresso server⁵⁹ within T-coffee⁶⁰, and this alignment was used to generate a neighbour joining tree in Geneious v.8.1.7.

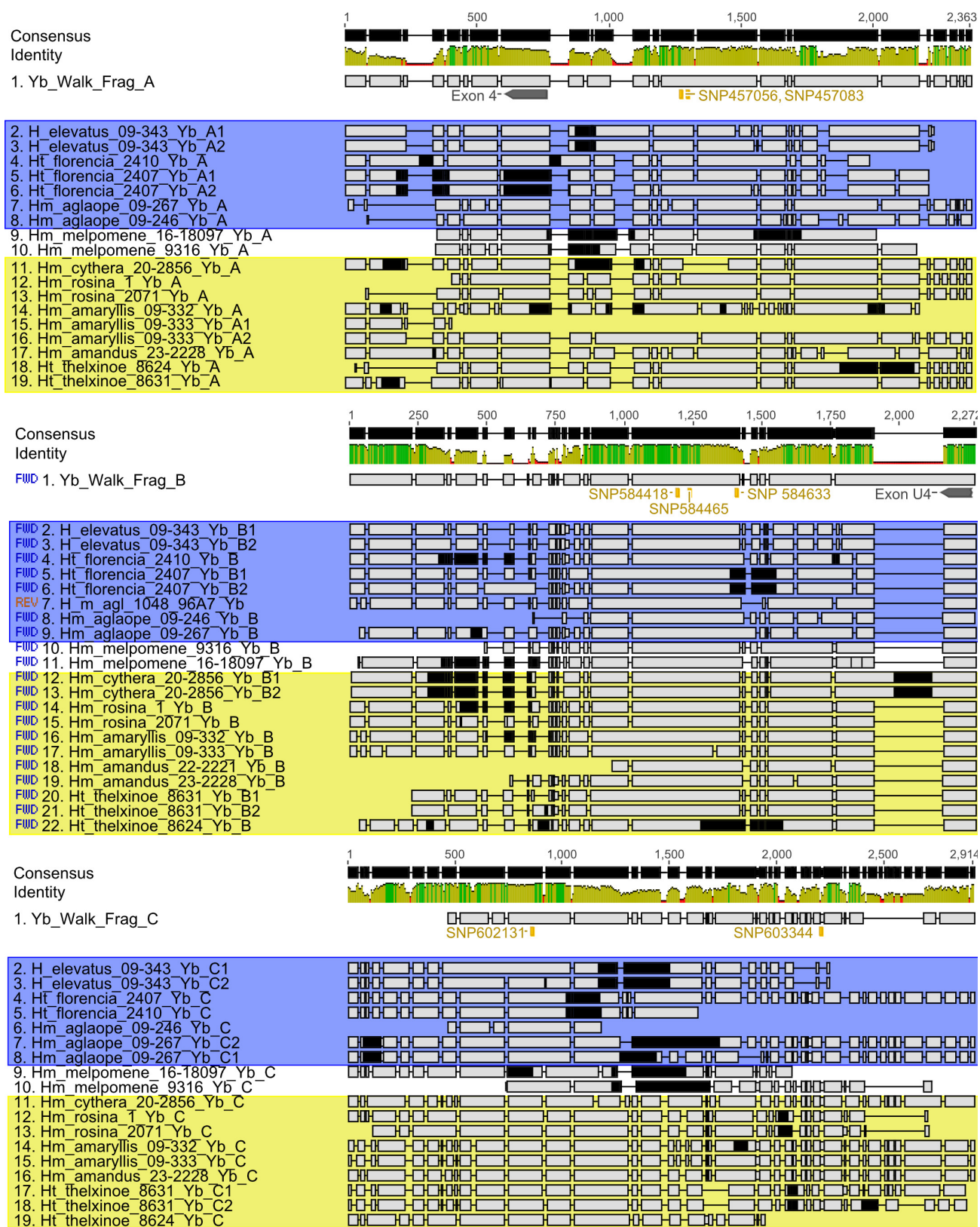
Expression of *H. melpomene cortex* in *D. melanogaster* wings. *D. melanogaster cortex* is known to generate an irregular microchaete phenotype when ectopically expressed in the posterior compartment of the adult fly wing²⁴. We performed the same assay using *H. melpomene cortex* to test whether this functionality was conserved. Following the methods of Swan and Schüpbach²⁴, we created an upstream activating sequence (UAS)–GAL4 construct using the coding region for the long isoform of *H. melpomene cortex*, plus a *Drosophila cortex* version to act as positive control. The haemagglutinin (HA)-tagged *H. melpomene* UAS-cortex expression construct was generated using cDNA reverse transcribed (Revert-Aid, Thermo-Scientific) from RNA extracted (Qiagen RNeasy) from pre-ommochrome pupal wing material. An HA-tagged *D. melanogaster* UAS-cortex version was also constructed²⁴. Expression was driven by the *hsp70* promoter. Constructs were injected into φC31-attP40 flies (25709, Bloomington Stock Centre; Cambridge University Genetics Department, UK, fly injection service) by site-directed insertion into CII via an attB site in the construct. Homozygous transgenic flies were crossed with *w^y;en-GAL4;UAS-GFP* flies (gift from M. Landgraf laboratory, Cambridge University Zoology Department) to drive expression in the engrailed posterior domain of the wing, and adult offspring wings were photographed (Extended Data Fig. 6b–d). Expression of the construct was confirmed by immunohistochemistry (using the standard *Drosophila* protocol) against an HA tag inserted at the N terminus of the protein, using final instar larval wing discs with mouse anti-HA and goat anti-mouse alexa-fluor 568 secondary antibodies (Abcam), imaged by Leica SP5 confocal (Extended Data Fig. 6e).

31. Supple, M. A. *et al.* Genomic architecture of adaptive color pattern divergence and convergence in *Heliconius* butterflies. *Genome Res.* **23**, 1248–1257 (2013).
32. de la Bastide, M. & McCombie, W. R. Assembling genomic DNA sequences with PHRAP. *Curr. Protoc. Bioinformatics* **11**, 11.4 (2007).
33. Gordon, D., Abajian, C. & Green, P. Consed: a graphical tool for sequence finishing. *Genome Res.* **8**, 195–202 (1998).
34. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
35. Magoč, T. & Salzberg, S. L. FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics* **27**, 2957–2963 (2011).
36. Andrews, S. *FastQC* <http://www.bioinformatics.babraham.ac.uk/projects/fastqc/> (2011).
37. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10**, R25 (2009).
38. Trapnell, C., Pachter, L. & Salzberg, S. L. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25**, 1105–1111 (2009).
39. Trapnell, C. *et al.* Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnol.* **28**, 511–515 (2010).
40. Holt, C. & Yandell, M. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinformatics* **12**, 491 (2011).
41. Slater, G. S. & Birney, E. Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* **6**, 31 (2005).
42. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
43. Lunter, G. & Goodson, M. Stampy: A statistical algorithm for sensitive and fast mapping of Illumina sequence reads. *Genome Res.* **21**, 936–939 (2011).
44. DePristo, M. A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genet.* **43**, 491–498 (2011).
45. Nadeau, N. J. *et al.* Genomic islands of divergence in hybridizing *Heliconius* butterflies identified by large-scale targeted sequencing. *Phil. Trans. R. Soc. B* **367**, 343–353 (2012).
46. Martin, S. H. *et al.* Genome-wide evidence for speciation with gene flow in *Heliconius* butterflies. *Genome Res.* **23**, 1817–1828 (2013).
47. Aulchenko, Y. S., Ripke, S., Isaacs, A. & van Duijn, C. M. GenABEL: an R library for genome-wide association analysis. *Bioinformatics* **23**, 1294–1296 (2007).
48. Smyth, G. K. in *Bioinformatics and Computational Biology Solutions Using R and Bioconductor* (eds Gentleman, R., Carey, V. J., Huber, W., Irizarry, R. A. & Dudoit, S.) 397–420 (Springer, 2005).
49. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. B* **57**, 289–300 (1995).
50. R Development Core Team. *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, 2011).
51. Darling, A. C. E., Mau, B., Blattner, F. R. & Perna, N. T. Mauve: multiple alignment of conserved genomic sequence with rearrangements. *Genome Res.* **14**, 1394–1403 (2004).
52. Jurka, J. *et al.* Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet. Genome Res.* **110**, 462–467 (2005).
53. Lavoie, C. A., Platt, R. N., Novick, P. A., Counterman, B. A. & Ray, D. A. Transposable element evolution in *Heliconius* suggests genome diversity within Lepidoptera. *Mob. DNA* **4**, 21 (2013).
54. Li, B. & Dewey, C. N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* **12**, 323 (2011).
55. Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome Biol.* **11**, R106 (2010).
56. Hardcastle, T. J. & Kelly, K. A. baySeq: Empirical Bayesian methods for identifying differential expression in sequence count data. *BMC Bioinformatics* **11**, 422 (2010).
57. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
58. Anders, S., Pyl, P. T. & Huber, W. HTSeq—A Python framework to work with high-throughput sequencing data. *Bioinformatics* **31**, 166–169 (2015).
59. Armougom, F. *et al.* Expresso: automatic incorporation of structural information in multiple sequence alignments using 3D-Coffee. *Nucleic Acids Res.* **34**, W604–W608 (2006).
60. Di Tommaso, P. *et al.* T-Coffee: a web server for the multiple sequence alignment of protein and RNA sequences using structural information and homology extension. *Nucleic Acids Res.* **39**, W13–17 (2011).



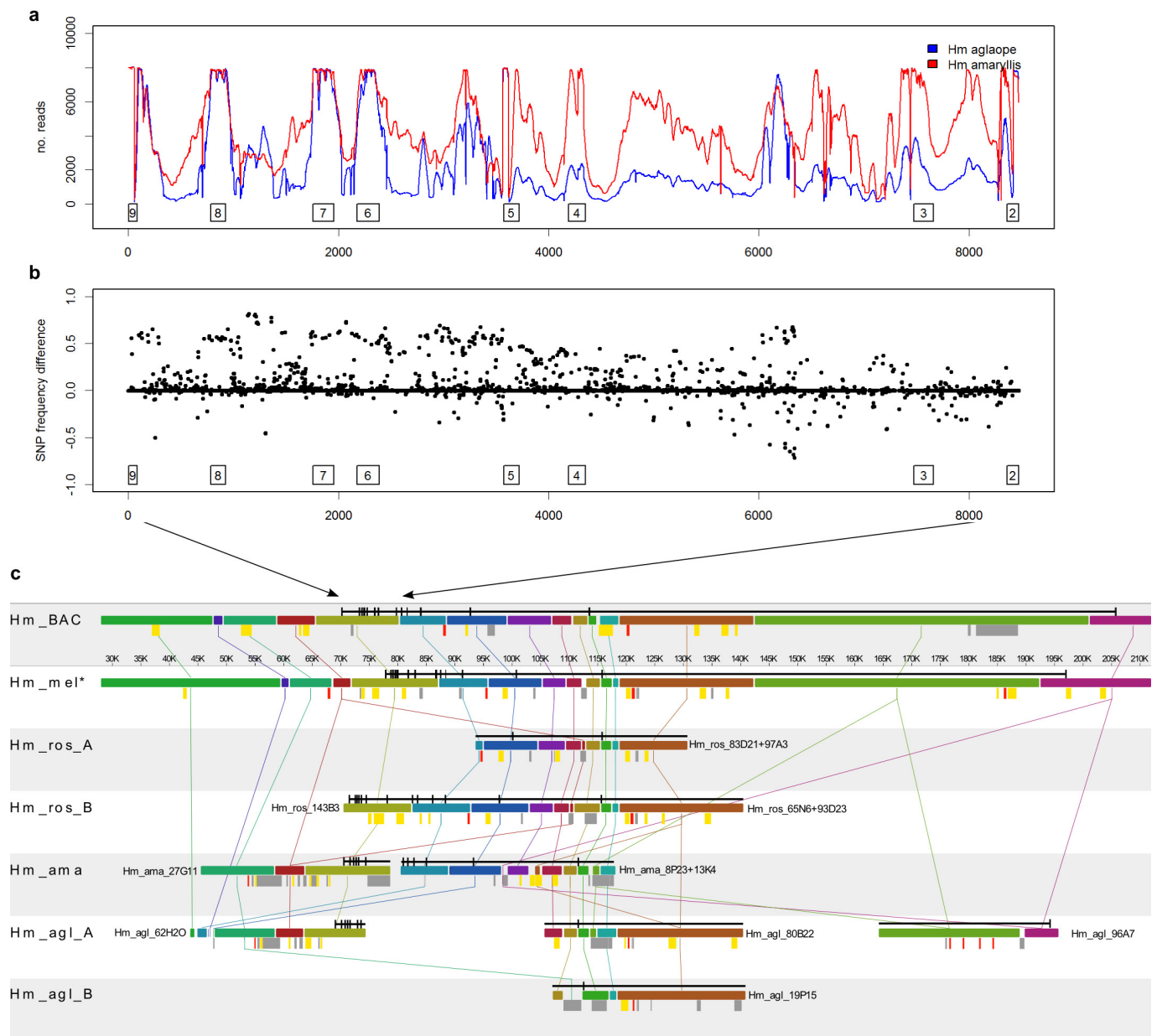
Extended Data Figure 1 | *H. melpomene* race-associated *cortex* splicing variation. **a**, Exons and splice variants of *cortex* in *H. melpomene*. Orientation is reversed with respect to Figs 2 and 4, with transcription going from left to right. SNPs showing the strongest associations with phenotype are shown with stars. **b**, Differential expression of two regions of *cortex* between whole hindwings of *H. melpomene amaryllis* and *H. melpomene aglaope* ($n = 11$ and $n = 10$, respectively). Box plots are standard (median; seventy-fifth and twenty-fifth percentiles; maximum and minimum excluding outliers (shown as discrete points)).

*** $P < 0.0001$, * $P < 0.05$, Wilcoxon rank sum test. **c**, Expression of a *cortex* isoform lacking exon 3 is found in *H. melpomene aglaope* but not *H. melpomene amaryllis* hindwings. **d**, Expression of an isoform lacking exon 5 is found in *H. melpomene rosina* but not *H. melpomene melpomene* hindwings. Green triangles indicate predicted start codons and red triangles predicted stop codons, with usage dependent on which exons are present in the isoform. Schematics of the targeted exons are shown for each (q)RT-PCR product; black triangles indicate the positions of the primers used in the assay.



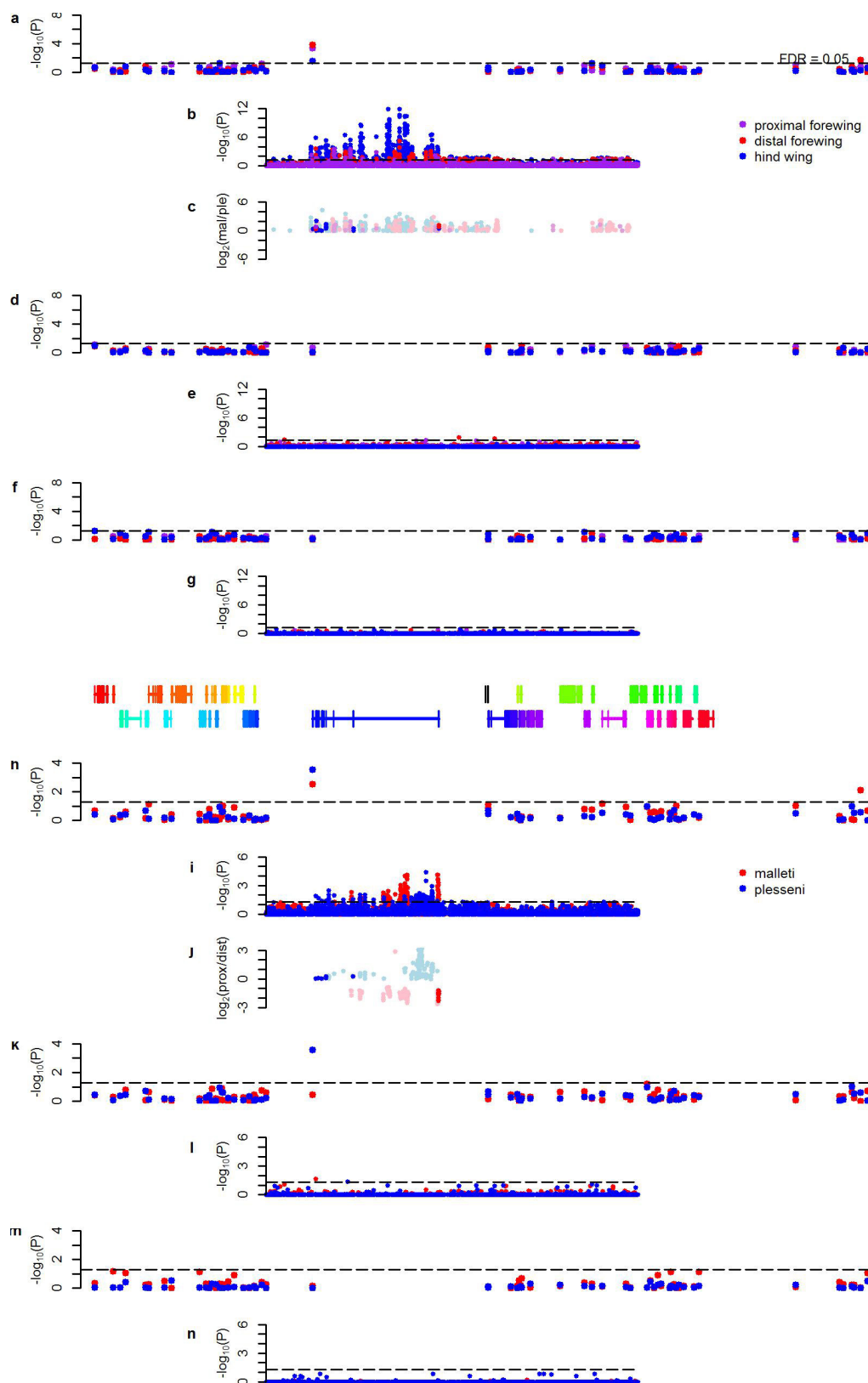
Extended Data Figure 2 | Alignments of *de novo* assembled fragments containing the top associated SNPs from *H. melpomene* and related taxa short-read data. Identified indels do not show stronger associations with phenotype than those seen at SNPs (as shown in Extended Data Table 2),

although some near-perfect associations are seen in fragment C. Black regions, missing data; yellow boxes, individuals with a yellow hindwing bar; blue boxes, individuals with a yellow forewing band.



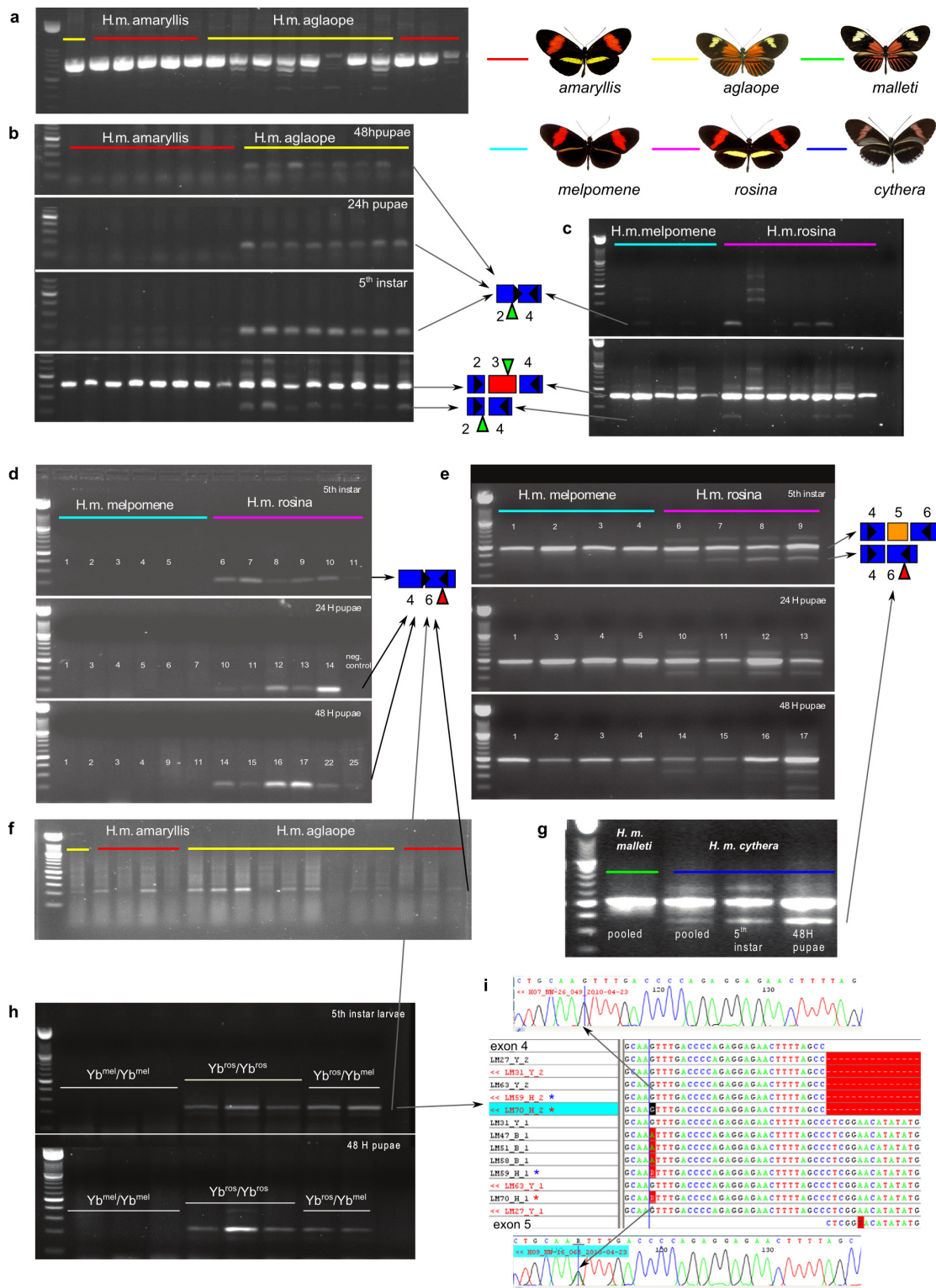
Extended Data Figure 3 | Sequencing of long-range PCR products and fosmids spanning *cortex*. **a**, Sequence read coverage from long-range PCR products across the *cortex* coding region from two *H. melpomene* races. **b**, Minor allele frequency difference from these reads between *H. melpomene aglaope* and *H. melpomene amaryllis*. Exons of *cortex* are indicated by boxes, numbered as in Extended Data Fig. 2. **c**, Alignments of sequenced fosmids overlapping *cortex* from three *H. melpomene* (*H. m.*) individuals of difference races. No major rearrangements are observed, nor any major differences in transposable element (TE) content between closely related races with different colour patterns (*melpomene/rosina* or *amaryllis/aglaope*). *H. melpomene amaryllis* and *rosina* have the same phenotype, but do not share any transposable elements that are

not present in the other races. Hm_BAC, BAC reference sequence; Hm_mel, *melpomene* from new unpublished assembly of *H. melpomene* genome⁵¹; Hm_ros, *rosina* (two different alleles were sequenced from this individual); Hm_ama, *amaryllis* (two non-overlapping clones were sequenced from this individual); Hm_agla, *aglaope* (four clones were sequenced from this individual, of which two represent alternative alleles). Alignments were performed with Mauve; coloured bars represent homologous genomic regions. *cortex* is annotated in black above each clone. Variable transposable elements are shown as coloured bars below each clone: red, Metulj-like non-LTR; yellow, Helitron-like DNA; grey, other.



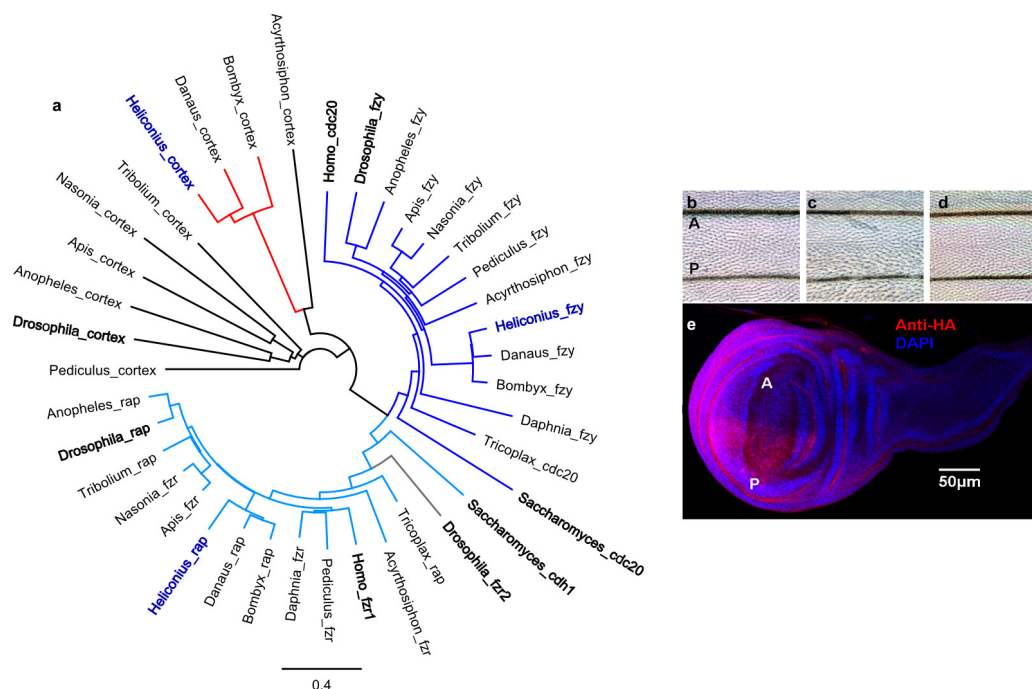
Extended Data Figure 4 | Expression array results for additional stages. Array results are related to Fig. 4. **a–g**, Comparisons between races (*H. melpomene plesseni* and *H. melpomene malleti*) for three wing regions. **h–n**, Comparisons between proximal and distal forewing regions for each race. Significance values ($-\log_{10}P$) are shown separately for genes in the *HmYb* region from the gene array (**a, d, f, h, k, m**) and for the *HmYb* tiling

array (**b, e, g, i, l, n**) for day 1 (**a, b, h, i**), day 5 (**d, e, k, l**) and day 7 (**f, g, m, n**) after pupation. The level of expression difference (log fold change) for tiling probes showing significant differences ($P \leq 0.05$) is shown for day 1 (**c** and **j**) with probes in known *cortex* exons shown in dark colours and probes elsewhere shown as pale colours. *P* values are based on FDR-adjusted *t*-statistics.



Extended Data Figure 5 | Alternative splicing of *cortex*. **a**, Amplification of the whole *cortex* coding region, showing the diversity of isoforms and variation between individuals. **b**, Differences in splicing of exon 3 between *H. melpomene aglaope* and *H. melpomene amaryllis*. Products amplified with a primer spanning the exon 2–4 junction at three developmental stages. The lower panel shows verification of this assay by amplification between exons 2 and 4 for the same final instar larval samples (replicated in Extended Data Fig. 2c). **c**, Lack of consistent differences between *H. melpomene melpomene* and *H. melpomene rosina* in splicing of exon 3. Top panel shows products amplified with a primer spanning the exon 2–4 junction; lower panel shows the same samples amplified between exons 2 and 4. **d**, Differences in splicing of exon 5 between *H. melpomene melpomene* and *H. melpomene rosina*. Products amplified with a primer spanning the exon 4–6 junction at three developmental stages. **e**, Subset of samples from **d** amplified with

primers between exons 4 and 6 for verification (middle, 24-h pupae samples are replicated in Extended Data Fig. 2d). **f**, Lack of consistent differences between *H. melpomene aglaope* and *H. melpomene amaryllis* in splicing of exon 5. Products amplified with a primer spanning the exon 4–6 junction. **g**, *H. melpomene cythera* also expresses the isoform lacking exon 5, while a pool of six *H. melpomene malleti* individuals do not. **h**, Expression of the isoform lacking exon 5 from an F₂ *H. melpomene melpomene* × *H. melpomene rosina* cross. Individuals homozygous or heterozygous for the *H. melpomene rosina* *HmYb* allele express the isoform while those homozygous for the *H. melpomene melpomene* *HmYb* allele do not. **i**, Allele-specific expression of isoforms with and without exon 5. Heterozygous individuals (indicated with blue and red stars) express only the *H. melpomene rosina* allele in the isoform lacking exon 5 (G at highlighted position), while they express both alleles in the isoform containing exon 5 (G/A at this position).



Extended Data Figure 6 | Phylogeny of fizzy family proteins and effects of expressing *cortex* in the *Drosophila* wing. **a**, Neighbour joining phylogeny of fizzy family proteins including functionally characterized proteins (in bold) from *Saccharomyces cerevisiae*, *Homo sapiens* and *D. melanogaster* as well as copies from the basal metazoan *Trichoplax adhaerens* and a range of annotated arthropod genomes (*Daphnia pulex*, *Acyrtosiphon pisum*, *Pediculus humanus*, *Apis mellifera*, *Nasonia vitripennis*, *Anopheles gambiae* and *Tribolium castaneum*) including the lepidoptera *H. melpomene* (in blue), *D. plexippus* and *B. mori*. Branch

colours: dark blue, cdc20/fzy; light blue, rap; red, lepidopteran cortex. **b–e**, Ectopic expression of *cortex* in *D. melanogaster*. *Drosophila cortex* produces an irregular microchaete phenotype when expressed in the posterior compartment of the fly wing (**c**) whereas *Heliconius cortex* does not (**d**), when compared to no expression (**b**). A, anterior; P, posterior. Successful *Heliconius cortex* expression was confirmed by anti-HA immunohistochemistry in the last instar *Drosophila* larva wing imaginal disc (**e**, red), with DAPI staining in blue.

Extended Data Table 1 | Genes in the *Yb* region and evidence for wing patterning control in *Heliconius*

<i>Hm</i> gene ID	<i>He</i> gene ID	Putative gene name	<i>Heliconius melpomene</i>										<i>H. erato</i>			<i>Hn</i>	
			Yb ^l	Sb ^l	A ^{Yb}	A ^N	E ¹	E ^{gw}	E ^{gr}	E ^{tw}	E ^{tr}	Cr ^l	A ^{pet}	A ^{fav}	P ^l	A ^{bic}	
HM00002	HERA000036	Acylpeptide hydrolase			2							x					
HM00003	HERA000037	HM00003										x					
HM00004	HERA000038	Trehalase-1B	x									x					
HM00006	HERA000038.1	Trehalase-1A	x									x					
HM00007	HERA000039	B9 protein	x									x					
HM00008	HERA000040	HM00008	x		2							x					
HM00010	HERA000041	WD40 repeat domain 85	x									x					
HM00012	HERA000042	CG2519	x					x				x					
HM00013	HERA000045	Unkempt	x									x					
HM00014	HERA000046	Histone H3	x									x					
HM00015	HERA000047	HM00015	x									x					
HM00016	HERA000048	HM00016	x									x					
HM00017	HERA000049	RecQ Helicase	x									x					
HM00018	HERA000051	HM00018	x									x					
HM00019	HERA000052	BmSuc2	x					x				x					
HM00020	HERA000053	CG5796	x									x					
HM00021	HERA000054	HM00021	x									x					
HM00022	HERA000055	Enoyl-CoA hydratase	x									x					
HM00023	HERA000056	ATP binding protein	x									x					
HM00024	HERA000057	HM00024	x									x					
HM00025	HERA000059	cortex	x	x	56	74	x	x	x	603	1796	x	2	99	x	51	
HM00026	HERA000077	Poly(A)-specific ribonuclease (parrn)		x	10					1	34	x				x	
HM00027	HERA000079	CG31320		x								x				x	
HM00028	HERA000080	ARP-like		x								x				x	
HM00029	HERA000081	CG4692		x								x				x	
HM00030	HERA000082	Proteasome 26S non ATPase subunit 4		x								x				x	
HM00031	HERA000083	HM00031		x					x			x				x	
HM00032	HERA000084	Zinc phosphodiesterase		x							1	x				x	
HM00033	HERA000085	Serine/threonine-protein kinase (LMTK1)		x							8	x				x	
HM00034	HERA000086	WD repeat domain 13 (Wdr13)			1	4					5	x				x	
HM00035	HERA000087	Domeless			1	2						x				x	
HM00036	HERA000061	WAS protein family homologue 1			5	36					37	x				x	
HM00038	HERA000062	Lethal (2) k05819 CG3054										x	2			x	
HM00039	HERA000064	Mitogen-activated protein kinase (MAPKK)										x				x	
HM00040	HERA000064.1	DNA excision repair protein ERCC-6										x				x	
HM00041	HERA000065	Penguin										x				x	
HM00042	HERA000066	Thymidylate kinase										x				x	
HM00043	HERA000067	Caspase-activated DNase										x				x	
HM00044	HERA000068	Regulator of ribosome biosynthesis										x				x	
HM00045	HERA000069	CG12659										x				x	
HM00046	HERA000070	CG33505										x				x	
HM00047	HERA000071	Sr protein										x				x	
HM00048	HERA000073	HM00048										x				x	
HM00049	HERA000073.1	HM00049										x				x	
HM00050	HERA000074	Shuttle craft										x				x	
HM00051	HERA000075	HM00051										x				x	
HM00052	HERA000076	HM00052					x					x				x	

A^{bic}, number of above background SNPs associated with the *H. numata* (*Hn*) *bicoloratus* phenotype in this study. A^{fav}, number of SNPs fixed for the alternative allele in *H. erato favorinus*. A^N, number of above background SNPs associated with the forewing yellow band in this study. A^{pet}, number of SNPs fixed for the alternative allele in *H. erato demophoon*. A^b, number of above background SNPs associated with the hindwing yellow bar in this study. Cr¹, within the previously mapped *HeCr* interval¹¹. P¹, within the previously mapped P interval¹³. E¹, detected as differentially expressed between *H. melpomene aglaope* and *amaryllis* from RNA-seq data in this study (Supplementary Information). E^{gr}, detected as differentially expressed between *H. melpomene plesseni* and *malleti* in the gene array in this study. E^{gw}, detected as differentially expressed between forewing regions in the gene array in this study. E^{lr}, numbers of probes showing differential expression between *H. melpomene plesseni* and *malleti* in the tiling array in this study. E^{tw}, numbers of probes showing differential expression between forewing regions in the tiling array in this study. Sb¹, within the previously mapped Sb interval¹². Yb¹, within the previously mapped Yb interval¹². Sb controls a white–yellow hindwing margin and is not investigated in this study. The N locus has not been fine-mapped previously.

Extended Data Table 2 | Locations of fixed or above-background SNPs and differentially expressed (DE) tiling array probes

		Positions of SNPs in the <i>He</i> and <i>Hn</i> association analyses								
		<i>cortex</i> coding exons	<i>cortex</i> UTR exons	<i>cortex</i> introns (nonTE)	<i>cortex</i> flanking intergenic (nonTE)	TEs	Other genes (exons or introns)	Other intergenic	Total	
<i>erato favorinus</i> fixed		2	0	96	8	2	0	0	108	
<i>erato demophoon</i> fixed		0	0	1	5	1	2	6	15	
<i>numata bicoloratus</i> above background		1	3	47	16	0	2	0	69	
Positions of DE tiling array probes		Known <i>cortex</i> coding exons	<i>cortex</i> UTR exons	<i>cortex</i> introns (nonTE)	miRNAs	TEs	Other gene exons	Other introns/intergenic	Total	
Day3	malleti vs plesseni	Forewing proximal	8	7	323	0	13	1	7	359
		Forewing distal	12	2	327	0	8	0	8	357
		Hindwing	5	14	378	0	9	1	6	413
	Proximal vs distal	malleti	0	1	68	0	0	0	12	81
		plesseni	2	4	222	0	10	0	4	242
Day1	malleti vs plesseni	Forewing proximal	1	0	22	0	3	0	7	33
		Forewing distal	2	3	116	1	9	5	112	248
		Hindwing	9	10	500	1	20	2	80	622
	Proximal vs distal	malleti	0	12	95	0	1	0	0	108
		plesseni	3	3	81	0	99	0	0	186

Extended Data Table 3 | SNPs showing the strongest phenotypic associations in the *H. melpomene/timareta/silvaniform* comparison

Species	Race	Sample Code	SNP pos HW 457083† bar (p=6.07E- 10)	SNP pos 439063* (p=1.72E- 09)	SNP pos 602131‡ (p=2.42E- 09)	SNP pos 457056† (p=2.42E- 09)	FW band	SNP pos 584465§ (p=1.37E- 07)	SNP pos 584418§ (p=1.41E- 07)	SNP pos 584633§ (p=2.10E- 07)	SNP pos 603344‡ (p=2.19E- 07)
<i>H. melpomene</i>	<i>aglaope</i>	09-246	0 A/A	A/G	A/A	C/C	1	T/T	A/A	NA	T/T
<i>H. melpomene</i>	<i>aglaope</i>	09-267	0 A/A	G/G	A/A	C/C	1	T/T	A/A	C/C	T/T
<i>H. melpomene</i>	<i>aglaope</i>	09-268	0 A/A	G/G	A/A	C/C	1	T/T	A/A	C/C	T/T
<i>H. melpomene</i>	<i>aglaope</i>	09-357	0 A/A	G/G	G/A	C/C	1	T/T	NA	C/C	T/T
<i>H. melpomene</i>	<i>aglaope</i>	aglaope.1	0 A/A	G/G	NA	C/C	1	C/T	T/A	T/C	T/T
<i>H. melpomene</i>	<i>amandus</i>	2221	1 A/A	NA	G/G	C/C	0	C/C	T/T	T/T	A/A
<i>H. melpomene</i>	<i>amandus</i>	2228	1 A/A	NA	G/G	C/C	0	C/T	T/A	T/C	A/A
<i>H. melpomene</i>	<i>amaryllis</i>	09-332	1 T/T	A/A	G/G	T/T	0	C/C	T/T	T/T	A/A
<i>H. melpomene</i>	<i>amaryllis</i>	09-333	1 T/T	A/A	G/G	T/T	0	C/C	T/T	T/T	A/A
<i>H. melpomene</i>	<i>amaryllis</i>	09-075	1 T/T	A/A	G/G	T/T	0	C/C	T/T	T/T	A/A
<i>H. melpomene</i>	<i>amaryllis</i>	09-079	1 T/T	A/A	G/G	T/T	0	C/C	T/T	T/T	A/A
<i>H. melpomene</i>	<i>amaryllis</i>	amaryllis.1	1 T/T	A/A	G/G	T/T	0	C/C	T/T	T/T	A/A
<i>H. melpomene</i>	<i>bellula</i>	228	1 T/T	NA	G/G	T/T	0	C/C	T/T	T/T	NA
<i>H. melpomene</i>	<i>bellula</i>	231	1 T/T	NA	G/A	T/T	0	C/T	T/A	T/C	NA
<i>H. melpomene</i>	<i>cythera</i>	2856	1 T/T	A/A	G/G	T/T	0	C/C	T/T	T/T	A/A
<i>H. melpomene</i>	<i>cythera</i>	2857	1 NA	NA	NA	NA	0	NA	NA	NA	NA
<i>H. melpomene</i>	<i>malleti</i>	17162	0 A/A	G/G	A/A	C/C	1	T/T	A/A	C/C	T/T
<i>H. melpomene</i>	<i>melpomene</i>	18038	0 A/A	G/G	G/G	C/C	0	C/C	T/T	T/T	A/A
<i>H. melpomene</i>	<i>melpomene</i>	18097	0 NA	G/G	NA	C/C	0	C/C	T/T	T/T	NA
<i>H. melpomene</i>	<i>melpomenem</i>	0.06	0 A/A	G/G	G/G	C/C	0	C/C	T/T	T/T	A/A
<i>H. melpomene</i>	<i>melpomenegen_ref</i>	0	0 A/A	G/G	NA	C/C	0	C/C	T/T	T/T	A/A
<i>H. melpomene</i>	<i>melpomene</i>	13435	0 A/A	G/G	A/A	C/C	0	C/C	T/T	T/T	A/A
<i>H. melpomene</i>	<i>melpomene</i>	9315	0 A/A	G/G	A/A	C/C	0	C/C	T/T	T/T	A/A
<i>H. melpomene</i>	<i>melpomene</i>	9316	0 A/A	G/G	A/A	C/C	0	C/C	T/T	T/T	A/A
<i>H. melpomene</i>	<i>melpomene</i>	9317	0 A/A	G/G	A/A	C/C	0	C/C	T/T	T/T	A/A
<i>H. melpomene</i>	<i>plesseni</i>	9156	0 A/A	G/G	A/A	C/C	0	C/C	T/T	T/T	NA
<i>H. melpomene</i>	<i>plesseni</i>	16293	0 A/A	G/G	A/A	C/C	0	C/C	T/T	T/T	NA
<i>H. melpomene</i>	<i>rosina</i>	rosina.1	1 T/T	A/A	G/G	T/T	0	C/C	T/T	T/T	A/A
<i>H. melpomene</i>	<i>rosina</i>	2071	1 T/T	A/A	G/G	T/T	0	C/C	T/T	T/T	A/A
<i>H. melpomene</i>	<i>rosina</i>	531	1 T/T	A/A	G/G	T/T	0	C/C	T/T	T/T	A/A
<i>H. melpomene</i>	<i>rosina</i>	533	1 T/T	NA	G/G	T/T	0	C/C	T/T	T/T	NA
<i>H. melpomene</i>	<i>rosina</i>	546	1 T/T	A/A	G/G	T/T	0	C/C	T/T	T/T	A/A
<i>H. melpomene</i>	<i>thelxiopiea</i>	13566	0 A/A	G/G	A/A	C/C	1	C/T	T/A	T/C	T/T
<i>H. melpomene</i>	<i>vulcanus</i>	14632	1 T/T	A/A	G/G	T/T	0	C/C	T/T	T/T	NA
<i>H. melpomene</i>	<i>vulcanus</i>	519	1 T/T	A/A	G/G	T/T	0	C/C	T/T	T/T	A/A
<i>H. timareta</i>	<i>florencia</i>	2403	0 A/A	G/G	A/A	C/C	1	T/T	A/A	C/C	T/T
<i>H. timareta</i>	<i>florencia</i>	2406	0 A/A	A/G	A/A	C/C	1	T/T	A/A	C/C	T/T
<i>H. timareta</i>	<i>florencia</i>	2407	0 A/A	A/G	A/A	C/C	1	T/T	A/A	C/C	T/T
<i>H. timareta</i>	<i>florencia</i>	2410	0 A/A	G/G	A/A	C/C	1	T/T	A/A	C/C	T/T
<i>H. timareta</i>	<i>timareta</i>	8533	0 A/A	G/G	A/A	C/C	1	C/T	T/A	T/C	T/T
<i>H. timareta</i>	<i>timareta</i>	9184	0 A/A	G/G	A/A	C/C	1	T/T	A/A	C/C	T/T
<i>H. timareta</i>	<i>timareta</i>	8520	0 A/A	G/G	A/A	C/C	1	T/T	A/A	C/C	T/T
<i>H. timareta</i>	<i>timareta</i>	8523	0 A/A	G/G	A/A	C/C	1	T/T	A/A	C/C	T/T
<i>H. timareta</i>	<i>thelxinoe</i>	09-312	1 T/T	A/A	G/G	T/T	0	C/C	T/T	T/T	A/A
<i>H. timareta</i>	<i>thelxinoe</i>	8624	1 T/T	A/A	G/G	T/T	0	C/C	T/T	T/T	A/A
<i>H. timareta</i>	<i>thelxinoe</i>	8628	1 T/T	A/A	G/G	T/T	0	C/C	T/T	T/T	A/A
<i>H. timareta</i>	<i>thelxinoe</i>	8631	1 T/T	A/A	G/G	T/T	0	C/C	T/T	T/T	A/A
<i>H. elevatus</i>		09-343	0 A/T	G/G	A/A	T/T	1	C/T	NA	C/C	T/T
<i>H. pardalinus</i>	<i>sergestus</i>	09-326	0 A/A	A/A	A/A	NA	0	C/C	T/T	T/T	NA

*Downstream of cortex. †Between exons 3 and 4 of cortex. ‡Upstream of cortex. §Between exons U4 and U3 of cortex. None of these SNPs are within known transposable elements. Colours show phenotypic associations: yellow, yellow hindwing bar; pink, no yellow hindwing bar; green, yellow forewing band; blue, no yellow forewing band; grey, allele does not match expected pattern.

Extended Data Table 4 | Transposable elements (TEs) found within the *Yb* region

Unique Occurrences					No.	TE name	Superfamily		Type
BAC	mel	ros	ama	agl					
1					1	BEL-1	BEL		LTR retrotransposon
	1				1	CR1-2	Jockey	LINE	Non-LTR retrotransposon
1					1	Daphne-1	Jockey	LINE	Non-LTR retrotransposon
1					1	Daphne-6	Jockey	LINE	Non-LTR retrotransposon
					1	DNA-like-8			DNA transposon
					1	Helitron-like-14	Helitron_A		DNA transposon
	1	2			4	Helitron-like-12	Helitron_A		DNA transposon
1	2				5	Helitron-like-12b	Helitron_A		DNA transposon
	1	1	1	1	7	Helitron-like-4a	Helitron_A		DNA transposon
						Helitron-like-4b	Helitron_A		DNA transposon
						Helitron-N2	Helitron_A		DNA transposon
					3	Helitron-like-7	Helitron_A		DNA transposon
5	3	3	1	2	16	Helitron-like-6a	Helitron_B		DNA transposon
						Helitron-like-6b	Helitron_B		DNA transposon
						Helitron-like-11	Helitron_B		DNA transposon
2	2	1		1	11	Helitron-like-15	Helitron_B		DNA transposon
6	5	3	1		18	Helitron-like-5	Helitron_B		DNA transposon
		1			2	Hmel_Unknown_50			
	1		1		2	Hmel_Unknown_174a/b			
	1				1	Hmel_Unknown_187b			
			1	1	2	Hmel_Unknown_230			
					1	Hmel_Unknown_234a			
					1	Hmel_Unknown_236a			
	1				1	Jockey-4	Jockey	LINE	Non-LTR retrotransposon
	1				1	LTR-3_gypsy	Gypsy		LTR retrotransposon
				1	1	Mariner-4	Mariner/Tc1		DNA transposon
1				3	29	Metulj-0	Metulj	SINE	Non-LTR retrotransposon
						Metulj-1	Metulj	SINE	Non-LTR retrotransposon
						Metulj-2	Metulj	SINE	Non-LTR retrotransposon
						Metulj-3	Metulj	SINE	Non-LTR retrotransposon
						Metulj-4	Metulj	SINE	Non-LTR retrotransposon
						Metulj-5	Metulj	SINE	Non-LTR retrotransposon
						Metulj-6	Metulj	SINE	Non-LTR retrotransposon
						Metulj-7	Metulj	SINE	Non-LTR retrotransposon
						nTc3-4	Mariner/Tc1		DNA transposon
						SINE-1	SINE	SINE	Non-LTR retrotransposon
1	1				2	nMar-3	Mariner/Tc1		DNA transposon
1					1	nMar-16	Mariner/Tc1		DNA transposon
			1		1	nMar-12/20	Mariner/Tc1		DNA transposon
				1	1	nPIF-3	PIF/Harbinger		DNA transposon
1					1	nTc3-2	Mariner/Tc1		DNA transposon
1					2	nTc3-3	Mariner/Tc1		DNA transposon
	1				2	R4-1	R2	LINE	Non-LTR retrotransposon
			1	1	6	Rep-1	REP	LINE	Non-LTR retrotransposon
2		1		1	4	RTE-3	RTE	LINE	Non-LTR retrotransposon
				1	2	RTE-11	RTE	LINE	Non-LTR retrotransposon
	1				3	Zenon-1	Jockey	LINE	Non-LTR retrotransposon
			1		1	Zenon-3	Jockey	LINE	Non-LTR retrotransposon

Early Neanderthal constructions deep in Bruniquel Cave in southwestern France

Jacques Jaubert^{1*}, Sophie Verheyden^{2,3*}, Dominique Genty^{4*}, Michel Soulier⁵, Hai Cheng^{6,7}, Dominique Blamart⁴, Christian Burtet², Hubert Camus⁸, Serge Delaby⁹, Damien Deldicque¹⁰, R. Lawrence Edwards⁷, Catherine Ferrier¹, François Lacrampe-Cuyaubère^{11,12}, François Lévêque¹³, Frédéric Maksud¹⁴, Pascal Mora¹⁵, Xavier Muth¹², Édouard Régner⁴, Jean-Noël Rouzaud¹⁰ & Frédéric Santos¹

Very little is known about Neanderthal cultures¹, particularly early ones. Other than lithic implements and exceptional bone tools², very few artefacts have been preserved. While those that do remain include red and black pigments³ and burial sites⁴, these indications of modernity are extremely sparse and few have been precisely dated, thus greatly limiting our knowledge of these predecessors of modern humans⁵. Here we report the dating of annular constructions made of broken stalagmites found deep in Bruniquel Cave in southwest France. The regular geometry of the stalagmite circles, the arrangement of broken stalagmites and several traces of fire demonstrate the anthropogenic origin of these constructions. Uranium-series dating of stalagmite regrowths on the structures and on burnt bone, combined with the dating of stalagmite tips in the structures, give a reliable and replicated age of 176.5 thousand years (± 2.1 thousand years), making these edifices among the oldest known well-dated constructions made by humans. Their presence at 336 metres from the entrance of the cave indicates that humans from this period had already mastered the underground environment, which can be considered a major step in human modernity.

Since its natural closing during the Pleistocene period and until its discovery⁶ in 1990, no humans entered Bruniquel Cave, located in southwest France (44° 4' N, 1° 41' E, Extended Data Fig. 1a), an area already rich in Palaeolithic sites (Extended Data Fig. 1b). Local cavers then dug through the collapsed entrance, a 30-m long and narrow passage through which persons can reach the main gallery. The structures (Fig. 1 and Extended Data Fig. 2a) are located at 336 m from the entrance after an easy walk through speleothem-rich chambers (Extended Data Fig. 1c). Near the entrance, the remains of large Pleistocene fauna and Holocene micro-fauna were found⁷, and bears also left numerous traces of their presence: hibernation hollows, claw marks and a few footprints. The most notable features, however, are the strange arrangement of two annular structures made of whole and broken stalagmites (Fig. 1 and Supplementary Video 1), accompanied by numerous traces of fire (Fig. 1 and Extended Data Fig. 3). Other than these structures, signs of human activity are almost non-existent and uncertain: a stalagmite tip that seems to have been hollowed out, negative prints left by wrenching stalagmites from the ground, and a few speleothem pieces in locations other than their original ones. At present, no marks on the cave walls or footprints have been observed. A first study in the early 1990s provided a detailed plan of the structures and a single ¹⁴C accelerator mass spectrometry dating of a burnt bone found in the main structure, giving an intriguing age of >47.6 thousand years ago (ka; ref. 6).

The question was whether these unique constructions were made by Neanderthals.^{8,9} Unfortunately, the premature death of the archaeologist F. Rouzaud, along with the restricted access to the cave, prevented any further research until 2013 when we decided to date and study these enigmatic constructions.

The arranged structures composed of whole and broken stalagmites, here designated as 'speleofacts' (Extended Data Table 1), are located in the largest chamber of the cave (Extended Data Fig. 1c). Our study defines two categories of structures: two annular ones, which are the most impressive, and four smaller stalagmite accumulation structures (Supplementary Video 1). The largest annular structure is 6.7×4.5 m, and the smaller one is 2.2×2.1 m. The accumulation structures consist of stacks of stalagmites and are from 0.55 m to 2.60 m in diameter. Two of them are located in the centre of the larger annular construction, while the other two are outside of it (Fig. 1). Overall, about 400 pieces were used, comprising a total length of 112.4 m and an average weight of 2.2 tons of calcite (Extended Data Table 1). Half of the elements composing the structures consist of the middle part of stalagmites (that is, without the root or tip), and very few pieces are whole ($\sim 5\%$). The stalagmites are well calibrated with a mean length of 34.4 cm for the large (A) and 29.5 cm for the small (B) annular structures, thus strongly suggesting intentional construction (Extended Data Fig. 4). Marks left by stalagmite wrenching are seen near the structures, though in most cases the original provenance of the stalagmites is difficult to determine owing to calcite flowstones covering a large part of the cave floor.

The annular structures are composed of one to four superposed layers of aligned stalagmites (Extended Data Fig. 2b). Notably, some short elements were placed inside the superposed layers to support them (Extended Data Fig. 2d, e). Other stalagmites were placed vertically against the main structure in the manner of stays, perhaps to reinforce the constructions (Extended Data Fig. 2a–c). All of these elements, combined with the large size of the structures, exclude any interventions by bears (Supplementary Information Table 2). Although bear traces are present throughout the cave (fur, claw marks, paw prints), hibernation hollows are observed only in other sectors (End Gallery, Bear Hollow Chamber at ~ 80 m and 240 m from the Structure Chamber).

Traces of fire are present on all six structures (Fig. 1). They consist of 57 reddened, more or less fissured speleofacts, and 66 blackened ones (Extended Data Fig. 3). The red and black colours are clearly not related to precipitates from the dripping water since no similar traces are observed on the ceiling. Instead, most of the coloured (and

¹PACEA, UMR 5199 CNRS-UB-MCC University of Bordeaux, 33615 Pessac, France. ²Earth & History of Life, Royal Belgian Institute of Natural Sciences, 1000 Brussels, Belgium. ³AMGC, Vrije Universiteit Brussel, 1050 Brussels, Belgium. ⁴LSCE, UMR 8212 CNRS-CEA-UVSQ, 91400 Gif-sur-Yvette, France. ⁵Société spéléologique et archéologique de Caussade, 5 rue Bourdelle 82300 Caussade, France. ⁶Institute of Global Environmental Change, Xi'an Jiaotong University, Xi'an 710049, China. ⁷Earth Sciences, University of Minnesota, Minneapolis, Minnesota 55455, USA. ⁸Protée Expert Sas, 30250 Sommières, France. ⁹Faculté Polytechnique, University of Mons, 7000-Mons, Belgium. ¹⁰Laboratoire de Géologie de l'École Normale Supérieure de Paris (ENS), UMR CNRS 8538, 75000 Paris, France. ¹¹Archéosphère, 11500 Quiribajou, France. ¹²Get in Situ, 1091 Bourg-en-Lavaux, Switzerland. ¹³LIENSs, UMR 7266 CNRS-University of La Rochelle, 17000 La Rochelle, France. ¹⁴Ministry of Culture, Regional Archaeological Service of Midi-Pyrénées, 31080 Toulouse, France. ¹⁵Archéostransfert, Archéovision, UMS 3657 SHS-3D, 33007 Pessac, France.

*These authors contributed equally to this work.

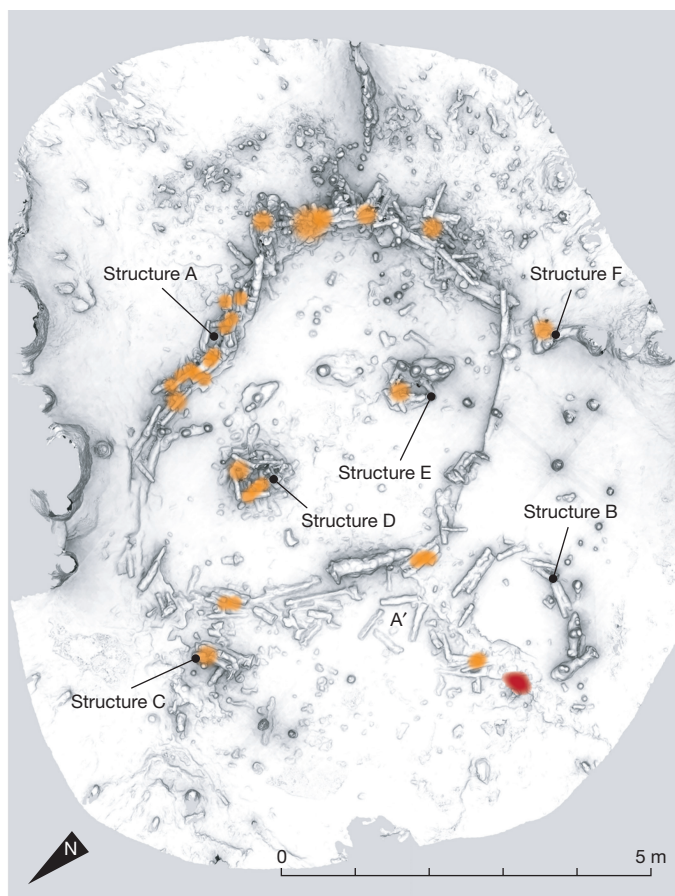


Figure 1 | Ortho-image of the Bruniquel Cave structures. The six structures are composed only of speleothems or fragments of speleothems (speleofacts), aligned and superimposed (A, B) (Extended Data Fig. 2a, b), or accumulated (C, D, E, F). A' is a likely extension of A. Their contours are sometimes imprecise due to the calcite layer and stalagmitic regrowths that cover them. The orange spots represent the heated zones, all located on the construction elements. The red spot (structure B) represents a char concentration (mainly burnt bone fragments) on the ground (Extended Data Fig. 3, bottom left).

fissured) locations were clearly heated, as confirmed by magnetic measurements of the most visibly reddened and blackened zones (Extended Data Fig. 5). A char (that is, carbonized organic material) is located near structure B, and a dozen black fragments are observed in the structures. The largest one is a 6.7-cm-long burnt bone (diaphysis) of a bear or large herbivore found on accumulation structure E (Fig. 1). It was covered by a 6-mm-thick calcite layer that has been precisely dated (Extended Data Table 2 and Extended Data Fig. 6a, d, e). The calcite surrounding this bone is reddened, blackened and fissured. Another black fragment was trapped between the calcite regrowth and the structure (Extended Data Fig. 6b). The black fragments and bone were clearly heated, as indicated by molecular and atomic spectrometry (Extended Data Fig. 6).

The age of the constructions has been determined by uranium-series dating of the stalagmite calcite (Supplementary Information SM2): the top of stalagmites that are part of the structure give maximum ages while the bases of the stalagmite regrowths sealing the structures give minimum ages (Supplementary Information SM1).

Eighteen multi-collector ICP-MS uranium-series ages¹⁰ with 2σ uncertainties were obtained from the calcite cores extracted from the stalagmites (Extended Data Table 2, Supplementary Information SM2).

Four additional samples were also dated: one from a core taken in the flowstone pavement inside annular structure A to evaluate its

contemporaneity with the structures, and three from the calcite layer that formed on the burnt bone found inside accumulation structure E (Fig. 1).

From the five calcite regrowths covering the structure, the two oldest ages are situated in the same time window, that is, 177.9 ± 3.7 ka and 175.2 ± 0.8 ka (Fig. 2 and Extended Data Table 2). They partially cover the age of the youngest dated stalagmite in the structure (177.1 ± 1.5 ka, Extended Data Table 2). All other ages correspond with this chronology, showing that the stalagmite tips are contemporary to or older than the calcite regrowths.

These results indicate that the structure was built between 175.2 ± 0.8 ka and 177.1 ± 1.5 ka (Fig. 3). Moreover, additional evidence for human presence in the cave at this time (Extended Data Table 2; Extended Data Figs 5 and 6) is provided by the burnt bone located in structure E, older than 180.9 ± 20.3 ka, the age of the calcite that formed on its surface, and the bone fragment trapped inside the BR-stm-SB7 core, with a minimum age of 175.2 ± 0.8 ka.

The age (175.9 ± 5.7 ka) of the calcite flowstone situated inside the annular structure is similar to that of the main structure within the margin of error (176.5 ± 2.1 ka), suggesting that the climate during this period (that is, 175–177 ka), covering part of marine isotope stage 6, was sufficiently humid and warm to allow continuous calcite deposition despite generally glacial conditions (Fig. 3). It can be associated with the warm phase VI-6-5 of the nearby Villars Cave speleothem record, characterized by low $\delta^{18}\text{O}$ and $\delta^{13}\text{C}$ (ref. 11) (Extended Data Fig. 7). Other European records show a similar climatic pattern, such as the high percentage of Euro-Siberian pollen in the MD01-2444 marine core off Lisbon between ~ 175 and 177 ka (ref. 11).

Early Neanderthals were the only human population living in Europe during this period¹². Our findings suggest that their society included elements of modernity, which can now be proven to have emerged earlier than previously thought. These include complex spatial organization, fire use, and deep karst occupation (Extended Data Fig. 8b).

Solid evidence for spatial organization (that is, human constructions, especially complex ones that required a social organization) during the Lower or Middle Palaeolithic is rare¹³. One hypothesis for its emergence postulates a sudden appearance of social organization with the arrival of modern humans (*Homo s. sapiens*)¹⁴, while a second hypothesis claims a more gradual and mosaic emergence during Neanderthal times in different parts of the world, including Europe¹⁵. In Europe, however, completely preserved sites are exceptional before the Upper Palaeolithic (42,000 calibrated years before present)¹⁶ and taphonomic processes hinder their identification^{17,18}. The spatial organization at Bruniquel Cave is the first one attributed with certitude to the early Middle Palaeolithic. The use of stalagmites is also unique for periods older than the Upper Palaeolithic, and implies a necessary simultaneous realization of different tasks and consequently, the existence of some degree of social organization (Extended Data Fig. 8a). The location of the Bruniquel structures inside a cave, where they were protected from weathering, animals and humans, played a major role in their preservation.

The first unequivocal use of fire is dated to the Middle Pleistocene (approximately 0.8 million years ago (Ma))¹⁹ and more than 1 Ma in southern Africa²⁰, with a more generalized use only after 0.3 Ma. A critical review of all known remains of fire in Europe²¹ concluded that Neanderthals were the first to commonly use fire, and in particular at the end of the Middle Pleistocene when they began to cook and produce new materials such as organic glue and haft tools. During marine isotope stage 6, the average number of proven fire uses for 10,000-year time slices is 1.47, which is very low²⁰. None of these sites is associated with a deep karst context.

Deep karst occupation does not appear to have occurred in Africa in any period, whether the Early or Middle Stone Age, or even the Late Stone Age if we exclude shelters and cave entrances with evidence for human presence in South Africa, Ethiopia and Maghreb (Extended

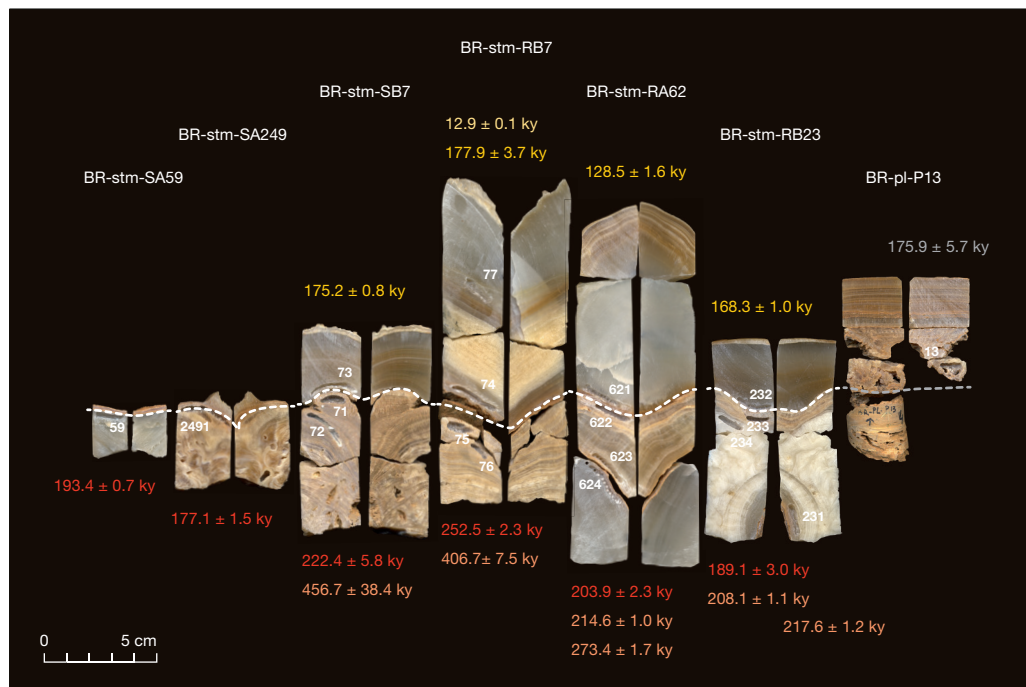


Figure 2 | The calcite cores sampled from the structures. BR-stm-SA59, BR-stm-SA249 and BR-stm-SB7 were cored from the tips of stalagmites used to build the structures. BR-stm-RB7, BR-stm-RA62 and BR-stm-RB23 were sampled at the base of stalagmites growing on the structures. All three cores display regrowth in their upper part as well as the older underlying stalagmite used as building item. Core BR-PL-P13 was taken from the flowstone located inside the main structure A. Samples were taken with a 1.6 cm (for BR-stm-SA59) and 2.6 cm diameter (for the other cores) coring device. Subsamples for uranium-series dating are indicated with their number (white). The dashed line indicates, within the

deposition of calcite, the moment of the building of the structures, that is, the limit between the stalagmites used in the structure (speleofact) and the regrowths. In most cases, this limit is marked by a clay layer. The ages for samples taken under the dashed line are given below the cores (orange); the ages for samples taken above the line (yellow) are given above the cores. The ages given in red are those which give the closest maximum age for the structures. ky, thousand years. The age of the flowstone inside the structure A is given in white, since the position corresponding to the time of construction (dashed line) inside the BR-PL-P13 core is still uncertain.

Data Fig. 8). The oldest evidence for the appropriation of this difficult environment is found in Europe²², Southeast Asia/Sunda²³, Wallacea²⁴ and Australia/Sahul²⁵. The accumulation of human bodies

by Acheuleans at Sima de los Huesos, Spain (0.35 Ma)²⁶ is very different from the Bruniquel structures, however. In other examples, the human frequentation of caves is linked to engraving, painting or

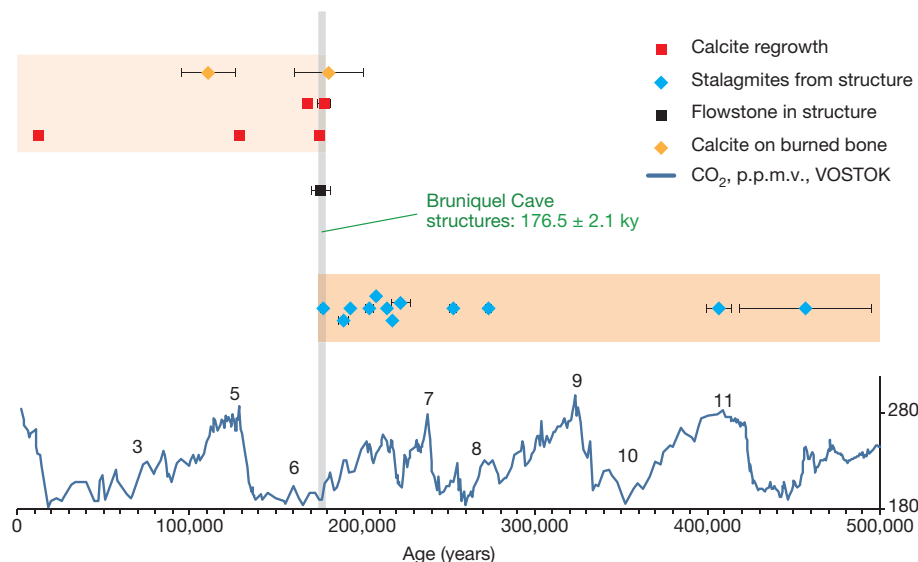


Figure 3 | Uranium-series ages (with 2σ error bars) obtained from the structures. Yellow, ages of the calcite covering the burnt bone in the accumulation structure E; red, ages obtained from the stalagmites covering the structure (regrowths) and representing a minimum age for the structure; blue, ages obtained from the stalagmites used by humans to build the structure (speleofacts) and representing a maximum age for the structures; black, age obtained from the flowstone partially covering

the inside area of the main structure. The age of the structures is situated between 175.2 ± 0.8 thousand years (ky) and 177.1 ± 1.5 ky. The calcite covering the burnt bone is dated to 180.9 ± 20.3 ka, indicating a minimum age of the bone and adding evidence of earlier human presence in the cave. The general climatic context is given by the CO₂ concentration variation (expressed in p.p.m.v., low right y axis) extracted from the Vostok ice core record³⁰ (black numbers indicate major marine isotope stages).

sculpting activities. These sites are thus younger than 42,000 calibrated years before present and are always associated with *Homo s. sapiens*. Symbolic, cultural or funerary activities were the main reasons for these cave visits. Until now no evidence has been found for regular Neanderthal incursions into caves, except for a possible case of footprints²⁷, and Neanderthal constructions inside caves, at least at a distance that is no longer exposed to daylight, were totally unknown. Moreover, Upper Palaeolithic constructions in caves are limited to fireplaces, simple hearths, and some rock or speleothem displacements. Even in caves regularly visited since the Aurignacian, constructions are non-existent or anecdotal^{28,29}.

What was the function of these structures at such a great distance from the cave entrance? Why are most of the fireplaces found on the structures rather than directly on the cave floor? Based on most Upper Palaeolithic cave incursions, we could assume that they represent some kind of symbolic or ritual behaviour³, but could they rather have served for an unknown domestic use or simply as a refuge? Future research will try to answer these questions.

The attribution of the Bruniquel constructions to early Neanderthals is unprecedented in two ways. First, it reveals the appropriation of a deep karst space (including lighting) by a pre-modern human species. Second, it concerns elaborate constructions that have never been reported before, made with hundreds of partially calibrated, broken stalagmites (speleofacts) that appear to have been deliberately moved and placed in their current locations, along with the presence of several intentionally heated zones. Our results therefore suggest that the Neanderthal group responsible for these constructions had a level of social organization that was more complex than previously thought for this hominid species.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 19 August 2015; accepted 27 April 2016.

Published online 25 May 2016.

1. Mellars, P. *The Neanderthal Legacy. An Archaeological Perspective from Western Europe*, (Princeton University Press, 1996).
2. Soressi, M. *et al.* Neandertals made the first specialized bone tools in Europe. *Proc. Natl Acad. Sci. USA* **110**, 14186–14190 (2013).
3. Soressi, M. & d'Errico, F. Pigments, gravures, parures: les comportements controversés des Néandertaliens. In *Les Néandertaliens, Biologie et Cultures* (eds Vandermeersch B. & Maureille B.) Doc. Préhist. **23**, Paris, CTHS, 297–309 (2007).
4. Maureille, B. & Vandermeersch, B. Les sépultures néandertaliennes. In *Les Néandertaliens, Biologie et Cultures* (eds Vandermeersch, B. & Maureille, B.) Doc. Préhist. **23**, Paris: CTHS, 311–322 (2007).
5. Villa, P. & Roebroeks, W. Neandertal Demise: An Archaeological Analysis of the Modern Human Superiority Complex. *PLoS One* **9**, e96424 (2014).
6. Rouzaud, F., Soulier, M. & Lignereux, Y. La grotte de Bruniquel. *Spelunca* **60**, 27–34 (1996).
7. Lafon, L. *La Grotte de Bruniquel (Tarn-et-Garonne). Inventaire au Sol des Vestiges Fauniques*. Thesis, University of Toulouse Paul Sabatier (1996).
8. Lorblanchet, M. *La Naissance de l'Art. Genèse de l'Art Préhistorique* (Paris, Errance, 1999).
9. Hayden, B. Neandertal social structure? *Oxf. J. Archaeol.* **31**, 1–26 (2012).
10. Cheng, H. *et al.* Improvements in ²³⁰Th dating, ²³⁰Th and ²³⁴U half-life values, and U-Th isotopic measurements by multi-collector inductively coupled plasma mass spectrometry. *Earth Planet. Sci. Lett.* **371–372**, 82–91 (2013).
11. Wainer, K. *et al.* Millennial climatic instability during penultimate glacial period recorded in a south-western France speleothem. *Palaeogeogr., Palaeoclim.* **376**, 122–131 (2013).
12. Hublin, J.-J. The origin of Neandertals. *Proc. Natl Acad. Sci. USA* **106**, 16022–16027 (2009).
13. Otte, M. The management of space during the Paleolithic. *Quatern. Int.* **247**, 212–229 (2012).

14. McBrearty, S. & Brooks, A. S. The revolution that wasn't: a new interpretation of the origin of modern human behavior. *J. Hum. Evol.* **39**, 453–563 (2000).
15. D'Errico, F. The invisible frontier. A multiple species model for the origin of behavioral modernity. *Evol. Anthropol.* **12–4**, 188–202 (2003).
16. Tchernych, A. P. in *Molodova I: A Single Case of Mousterian Settlement in the Middle Dniestr Basin* (in Russian) (ed. Goretsky, G.I. & Ivanova, I.K.) 6–102 (Nauka, 1982).
17. de Lumley, H. *Une Cabane Acheuléenne dans la Grotte du Lazaret à Nice*. Vol. 7, Paris, Soc. Préhist. Franç. (1969).
18. Mania, D. H. *et al.* *Bilzingsleben II. Homo erectus – Seine Kultur und Seine Umwelt*. VEB Deutscher Verlag der Wissenschaften, Berlin (1983).
19. Goren-Inbar, N. *et al.* Evidence of hominin control of fire at Gesher Benot Ya'aqov, Israel. *Science* **304**, 725–727 (2004).
20. Berna, F. *et al.* Microstratigraphic evidence of in situ fire in the Acheulean strata of Wonderwerk Cave, Northern Cape Province, South Africa. *Proc. Natl Acad. Sci. USA* **109**, 1215–1220 (2012).
21. Roebroeks, W. & Villa, P. On the earliest evidence for habitual use of fire in Europe. *Proc. Natl Acad. Sci. USA* **108**, 5209–5214 (2011).
22. Clottes, J. *et al.* Les peintures paléolithiques de la grotte Chauvet-Pont d'Arc, à Vallon-Pont-d'Arc (Ardèche, France): datations directes et indirectes par la méthode du radiocarbon. *C.R. Acad. Sc. Paris* **320**, 1133–1140 (1995).
23. Plagnes, V. *et al.* Cross dating (Th/U-¹⁴C) of calcite covering prehistoric paintings in Borneo. *Quat. Res.* **60**, 172–179 (2003).
24. Aubert, M. *et al.* Pleistocene cave art from Sulawesi, Indonesia. *Nature* **514**, 223–227 (2014).
25. Bednarik, R. The cave petroglyphs of Australia. *Aust. Aborig. Stud.* **2**, 64–68 (1990).
26. Arsuaga, J. L. Bermudez de Castro, J. M. & Carbonell, E. (Eds) The Sima de los Huesos hominid site. *J. Hum. Evol.* **33** (special issue).
27. Onac, B. P. *et al.* U-Th ages constraining the Neanderthal footprint at Vârtope Cave, Romania. *Quat. Sci.* **24**, 1151–1157 (2005).
28. Bégouën, R., Clottes, J., Feruglio, V. & Pastors, A. *La Caverne des Trois-Frères*, Paris Somogy- Assoc. L. Bégouën (2013).
29. Delannoy, J.-J. *et al.* The social construction of caves and rockshelters: Chauvet Cave (France) and Nawarla Gabarnmang (Australia). *Antiquity* **87**, 12–29 (2013).
30. Petit, J. R. *et al.* Climate and atmospheric history of the past 420,000 years from the Vostok ice core, Antarctica. *Nature* **399**, 429–436 (1999).

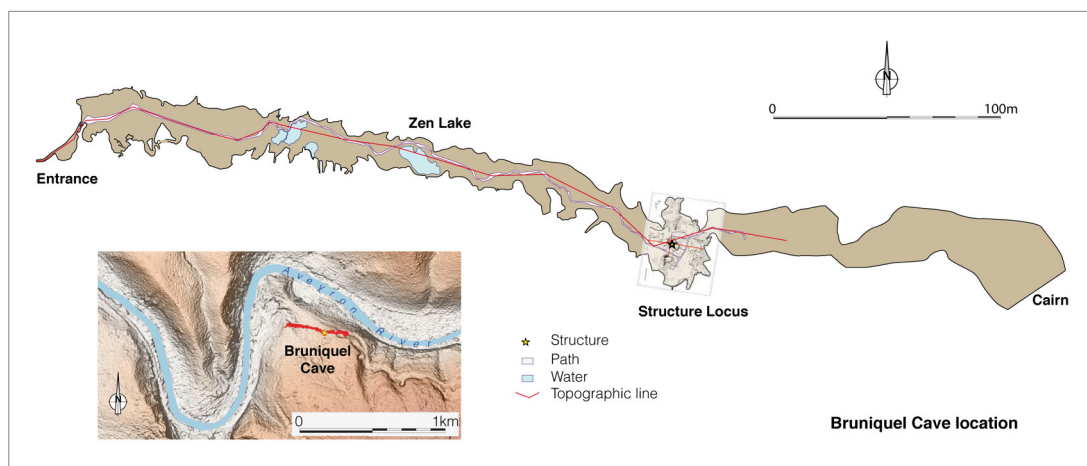
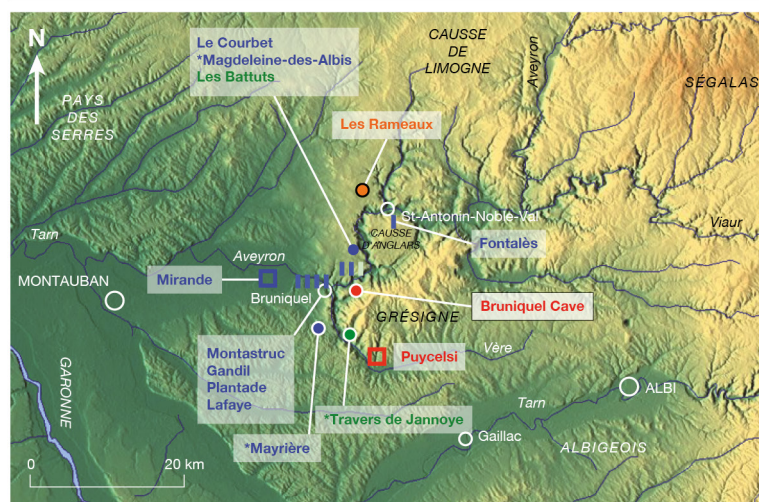
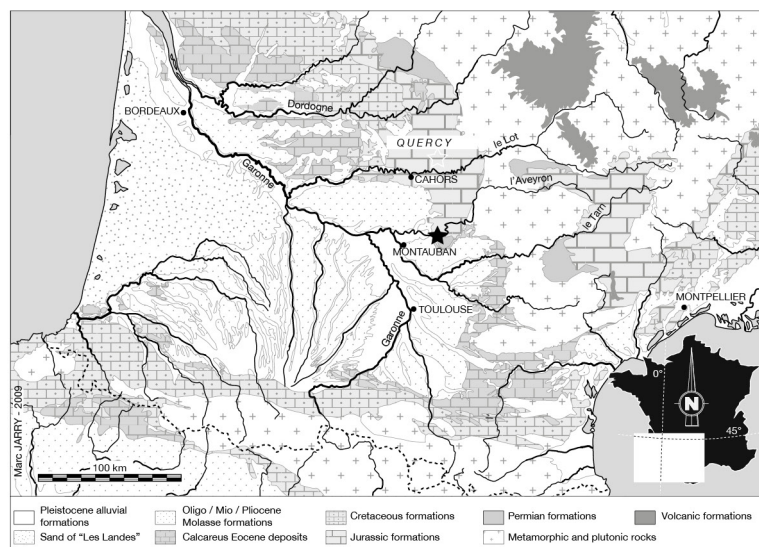
Supplementary Information is available in the online version of the paper.

Acknowledgements We thank the owners of the cave (Nidauzel association), the French Ministry of Culture & Communication, MCC (DRAC-SRA Midi-Pyrénées, Toulouse), M. Vaginay, P. Chalard, É. Mauduit, the Speleological & Archaeological Society of Caussade (SSAC), CNRS (InEE & InSU), the University of Bordeaux-PACEA, LSCE Gif-s/-Yvette, M. O'Farrell and C. Garrec for editing, V. Feruglio for a drawing. We thank F. Dewilde and F. Mansouri (LSCE) for their assistance with the isotopic measurements, Y. Vanbrabant (Belgian Geological Survey) for his assistance with the cave monitoring and B. Martinez for his help with the topography. We thank S. Mariot and R. Weil (LPS, Paris-XI University, Orsay) for their help in the infrared spectrometry measurements. This work is mainly supported by French MCC (DRAC-SRA Midi-Pyrénées, Toulouse) and in part by the Belgian Science Policy Office. The U-Th dating was supported in part by the U.S. NSF.

Author Contributions J.J., S.V. and D.G. coordinated this study; they wrote the article and conducted the field sampling. M.S. participated in the cave discovery and is in charge of the logistical support and cave access. H.Ch. made the U-Th measurements and R.L.E. oversaw and helped to interpret the U/Th dates. D.B. conducted the ^δ¹⁸O and ^δ¹³C measurements. C.B. is responsible for the temperature monitoring. H.C., S.D. and X.M. realised the geographical and new topography studies of the cave. F.L.-C. realised the drawings. F.L. realised the magnetism measurements and their interpretation, D.D., D.G. and J.-N.R. the SEM-EDS, FTIR measurements and Raman spectrometry. F.M. participated in the field trips and archaeological survey. P.M. realised the photogrammetric work. C.F. realised the study of fireplaces and heated areas. É.R. participated in the field trips and the coring. F.S. is responsible for the statistical studies of the structure elements.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to J.J. (jacques.jaubert@u-bordeaux.fr) or S.V. (sophie.verheyden@naturalsciences.be) or D.G. (dominique.genty@lsce.ipsl.fr).

31. Rouzard, F., Soulier, M., Brugal, J.-Ph. & Jaubert, J. L'igue des Rameaux (Saint-Antonin-Noble-Val, Tarn-et-Garonne). Un nouveau gisement du Pléistocène moyen. Premiers résultats. *Paléo* **2**, 89–106 (1990).
32. Faivre, J.-Ph. *et al.* In *Modalités d'Occupations et Exploitation des Milieux au Paléolithique dans le Sud-Ouest de la France: l'Exemple du Quercy* (eds Jarry, M., Brugal, J.-Ph. & Ferrier, C.), *Paléo* **4** (suppl.), 231–270 (2013).
33. Clottes, J. *et al.* In *L'Art des Cavernes. Atlas des Grottes Ornées Françaises* (Paris, Ministère de la Culture-Imprimerie Nationale), 540–551 (1984).
34. Cantalejo, P., del Mar Espejo, M., Ramos, J. & Weniger, G. C. Elementos de iluminación in *Cueva de Ardales. Intervenciones arqueológicas 2011-2014* (eds Ramos, J., Weniger, G.C., Cantalejo, P. & del Mar Espejo, M.) Ediciones Pinsapar, 119–146 (2014).
35. Brodard, A. *et al.* In *Actes du colloque MADAPCA, Micro Analyses et Datations de l'Art Préhistorique dans son Contexte Archéologique, MNHN-C2RMF, 16-18 novembre 2011. Paléo* (special issue) 233–235 (2014).
36. Bertran, P., Hétu, B. & Texier, J.-P. Fabric characteristics of subaerial slope deposits. *Sedimentology* **44**, 1–16 (1997).
37. Lenoble, A. & Bertran, P. Fabric of Palaeolithic levels: methods and implications for site formation processes. *J. Archaeol. Sci.* **31**, 457–469 (2004).
38. Scollar, I., Tabbagh, A., Hesse, A. & Herzog, I. *Archaeological Prospecting and Remote Sensing*. Cambridge, Cambridge University Press (1990).
39. Le Borgne, E. Influence du feu sur les propriétés magnétiques du sol et sur celles du schiste et du granite. *Ann. Geophys.* **16**, 159–195 (1960).
40. Maki, D., Homburg, J. A. & Brosowski, S. D. Thermally activated mineralogical transformations in archaeological hearths: inversion from maghemite (γ -Fe₂O₃) phase to hematite (α -Fe₂O₃) form. *Archaeol. Prospect.* **13**, 207–227 (2006).
41. Carrancho, Á. & Villalán, J. J. Different mechanisms of magnetisation recorded in experimental fires: archaeomagnetic implications. *Earth Planet. Sci. Lett.* **312**, 176–187 (2011).
42. Jrad, A. *et al.* Magnetic investigations of buried palaeohearths inside a Palaeolithic cave (Lazaret, Nice, France). *Archaeol. Prospect.* (2013).
43. Brodard, A. *et al.* Thermal characterization of ancient hearths from the cave of Les Fraux (Dordogne, France) by thermoluminescence and magnetic susceptibility measurements. *Quat. Geochronol.* **10**, 353–358 (2012).
44. Burens, A. *et al.* Benefits of an accurate 3D Documentation in Understanding the Status of the Bronze Age Heritage Cave 'Les Fraux' (France). *Int. J. of Heritage in the Digital Era* **1**, 179–195 (2014).
45. Stiner, M. C. & Kuhn, S. L. Differential burning, recrystallization, and fragmentation of archaeological bone. *J. Archaeol. Sci.* **22**, 223–237 (1995).
46. Lebon, M. *et al.* Application des micro-spectrométries infrarouges et Raman à l'étude des processus diagénétiques altérant les ossements paléolithiques. *Rev. Archéométrie* **35**, 179–190 (2011).
47. Ellingham, S. T. D., Thompson, T. J. U., Islam, M. & Taylor, G. Estimating temperature exposure of burnt bone - a methodological review. *Sci. Justice* **55**, 181–188 (2015).
48. Rouzard, J.-N., Deldicque, D., Charon, E. & Pageot, J. Carbons in the heart of energy and environment questions: a nanostructural approach. *C. R. Geosci.* **347**, 124–133 (2015).
49. Deldicque, D., Rouzard, J.-N. & Velde, B. A Raman-HRTEM study of the carbonization of wood: a new Raman-based paleothermometer dedicated to archaeometry. *Carbon* (2016).
50. Wainer, K. *et al.* Speleothem record of the last 180 ka in Villars cave (SW France): investigation of a large delta (δ^{18} O) shift between MIS6 and MIS5. *Quat. Sci. Rev.* **30**, 130–146 (2011).

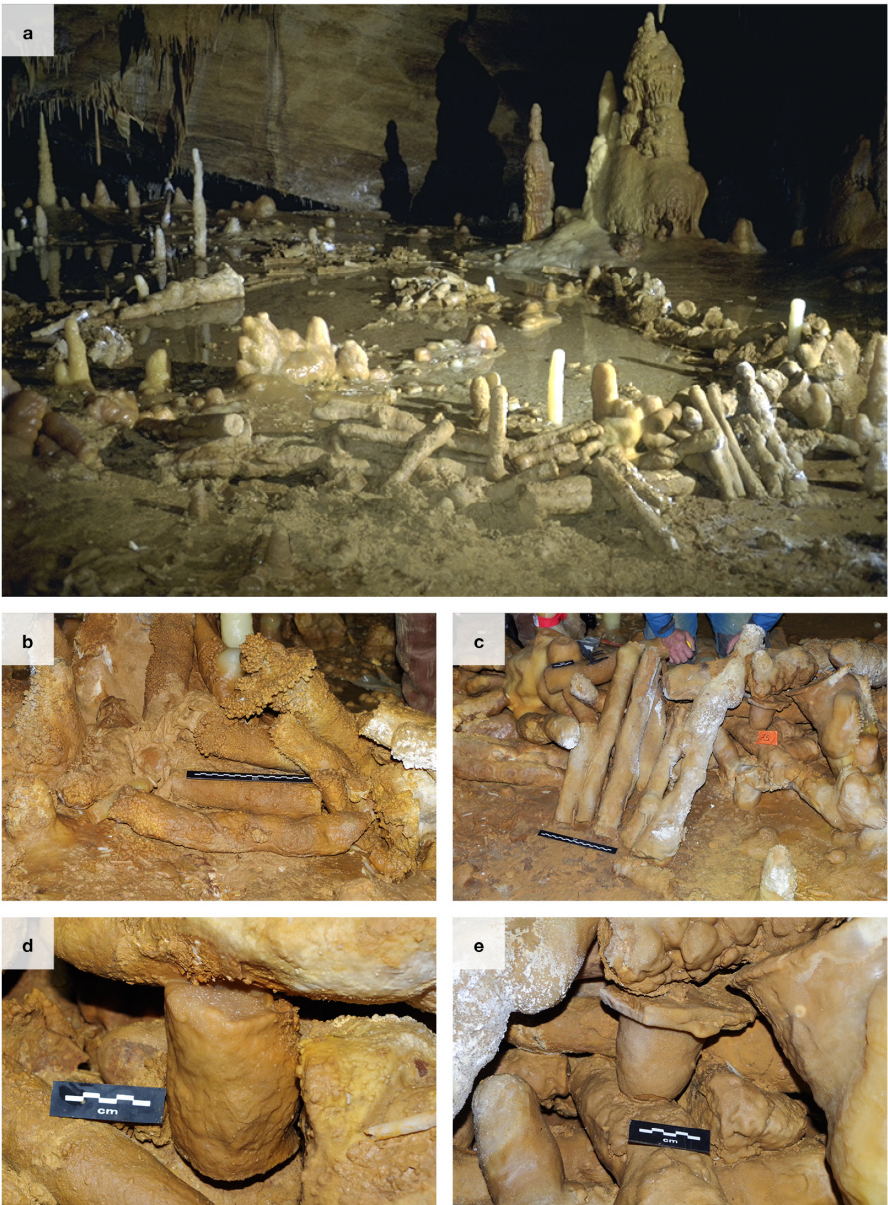


Extended Data Figure 1 | See next page for caption.

Extended Data Figure 1 | Location and map of Bruniquel Cave.

a, Bruniquel Cave (marked with a star) is located in the southwest of France, south of the calcareous plateaus of Quercy, east of the Aquitaine Basin. Its entrance (165 m above sea level) overlooks the Aveyron valley, a tributary of the Tarn on the right bank of the Garonne and down from the Massif Central (base map courtesy of M. Jarry). **b**, Bruniquel Cave in the Aveyron valley. Orange: Lower Palaeolithic site; red: Middle Palaeolithic sites; green: early Upper Palaeolithic; blue: late Upper Palaeolithic (Magdalenian). Circles indicate caves, vertical lines indicate rock shelters and squares mark open-air sites. *Decorated caves. In this area within a 30 km zone around Bruniquel Cave, fifteen major Palaeolithic sites are known. The oldest known human occupations in this region are those of the Igüe des Rameaux (Tarn-et-Garonne), a karstic sinkhole where lithic material was associated with a recent mid-Pleistocene fauna, dated from marine isotope stages 9 to 5 (ref. 31). A Middle Palaeolithic, stratified open-air site is also present at La

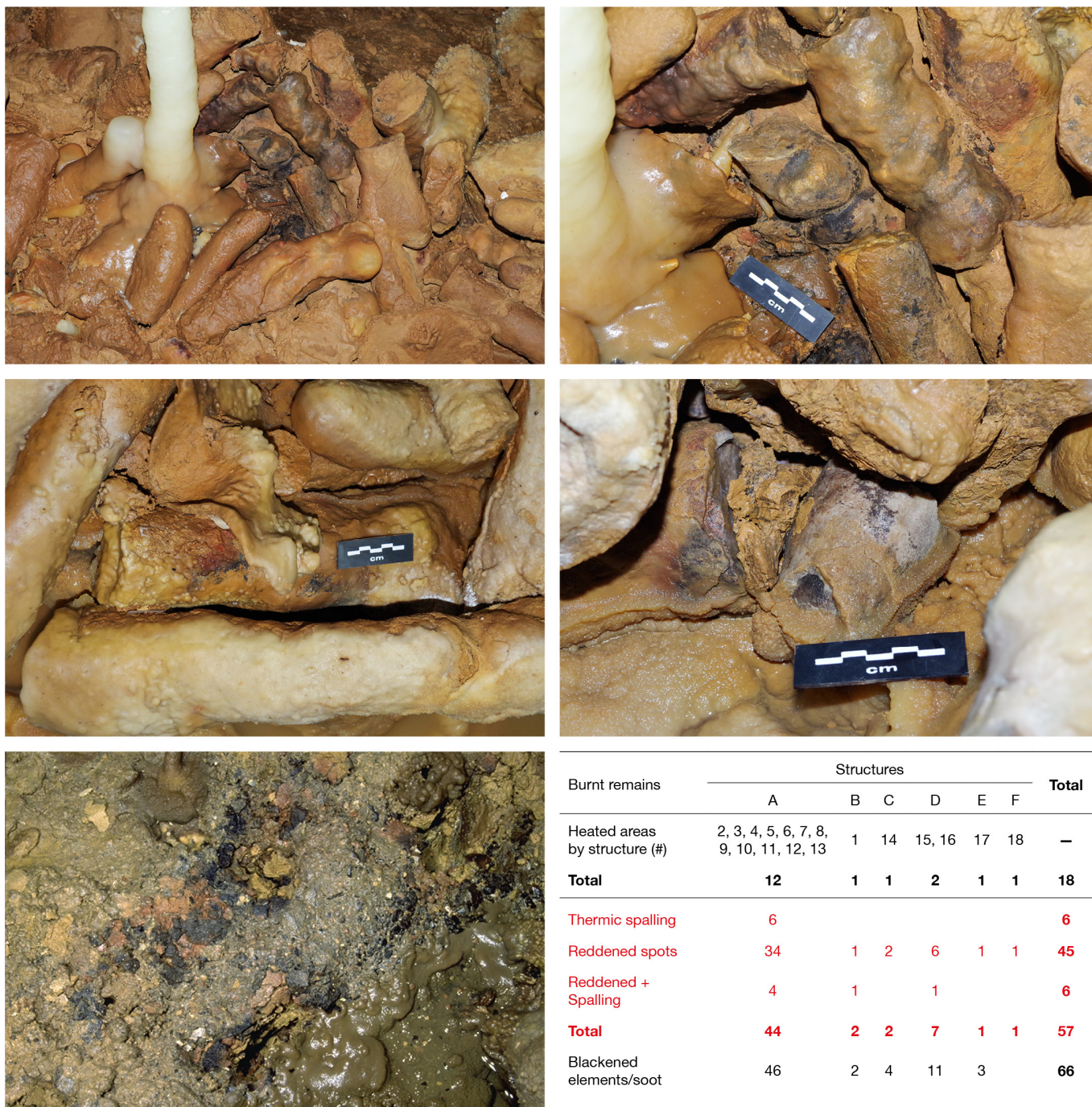
Rouquette-Puycelsi (Tarn) upstream on the nearby Vère River³². The other sites are all attributable to the Upper Palaeolithic, representing the Aurignacian, Gravettian and Solutrean periods, but mainly the Magdalenian period with three decorated caves: Travers de Jannoye, La Magdeleine-des-Albis (Penne, Tarn) and Mayrière (Bruniquel, Tarn-et-Garonne)³³ (base map, courtesy of StepMap GmbH, modified by J.J.). **c**, Topography of Bruniquel Cave. The cave consists of a 10–15 m wide and 4–7 m high corridor, currently known to be 482 m long. Beyond the narrow entrance passage (filled porch), there are no major topographic difficulties until the chamber containing the structure at 336 m from the unobstructed entrance. Currently, no other access has been identified, laterally or at the other end. In this latter case, a second obstructed entrance would be at least 295 m from another slope. Sources: Structure drawn by M. Soulier and F. Rouzaud, 1992; topography realized by Protée-Expert & Get in Situ, 2015; Digital Elevation Model generated with 1957 aerial photography IGN, public domain).



f	Speleofacts		Structures						Total
			A	B	C	D	E	F	
Structures	Total Number		267	49	9	53	15	6	399
	Maximum length (m)	internal	5.80	1.60	2.60	1.30	1.15	0.55	—
		external	6.70	2.20					
	Maximum width (m)	internal	3.70	1.50	0.90	1.15	0.85	0.50	—
		external	4.50	2.10					
	Circumference (m)	internal	16	5.45	0.60	4.05	3.65	1.60	37.95
		external	20.65	7.4					
Surface (m²)	internal	16.3	2.3	0.55	1.5	1	0.3	29.35	
	external	23	3						
Traces of fire		12	1	1	2	1	1	18	

Extended Data Figure 2 | Bruniquel Cave structures. **a**, General view of the main structure (structure A) with superposed layers of aligned stalagmites (speleofacts) Photo courtesy of É. Fabre, SSAC. **b**, Example of speleofacts accumulated over three or even four horizontal levels. **c**,

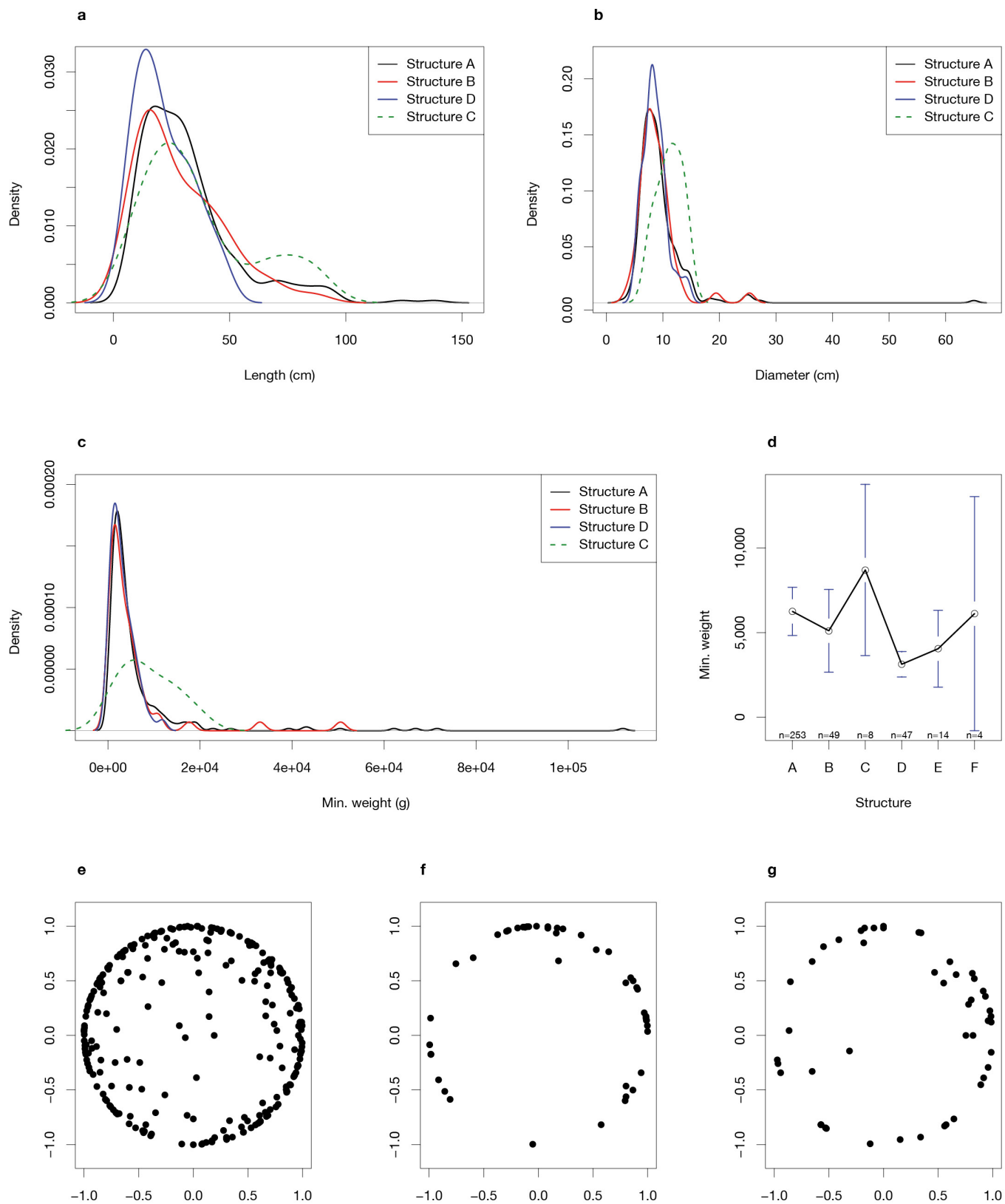
Stalagmites (speleofacts) placed vertically against the main structure (structure A) in the manner of stays. **d**, **e**, Two examples of short back stalagmites serving as sustaining pieces. **f**, Summary of the metric data of the structures.

**a**

Extended Data Figure 3 | Fireplaces and heated areas. **a**, Examples of a fireplace on the main structure. Note the reddened, blackened and fissured stalagmites³⁴. The structure in this location (top) is covered by white, more recent and still active stalagmites. The heated areas on the speleofacts correspond to the red and grey colours, as well as fissuring and superficial spalling. These scars are similar to thermal alterations studied in the cave

b

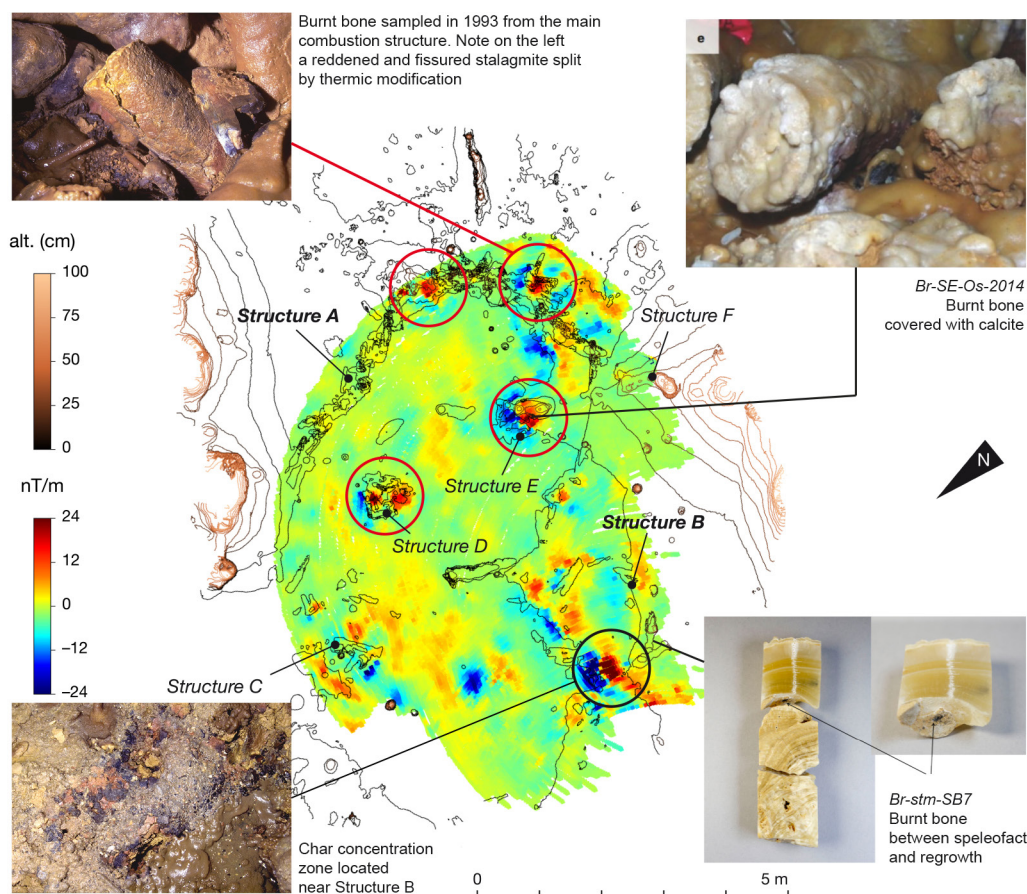
of Chauvet-Pont d'Arc (Ardèche)³⁵. In our current stage of observation, the study of their distribution enabled us to identify a well-preserved fireplace in structure A, as well as structures that have been disturbed by processes that remain to be determined (structures D and E, for example). **b**, Numbers per structure of heated areas, thermic spalling, fissured spots and blackened elements (that is, speleofacts) and soot.



Extended Data Figure 4 | See next page for caption.

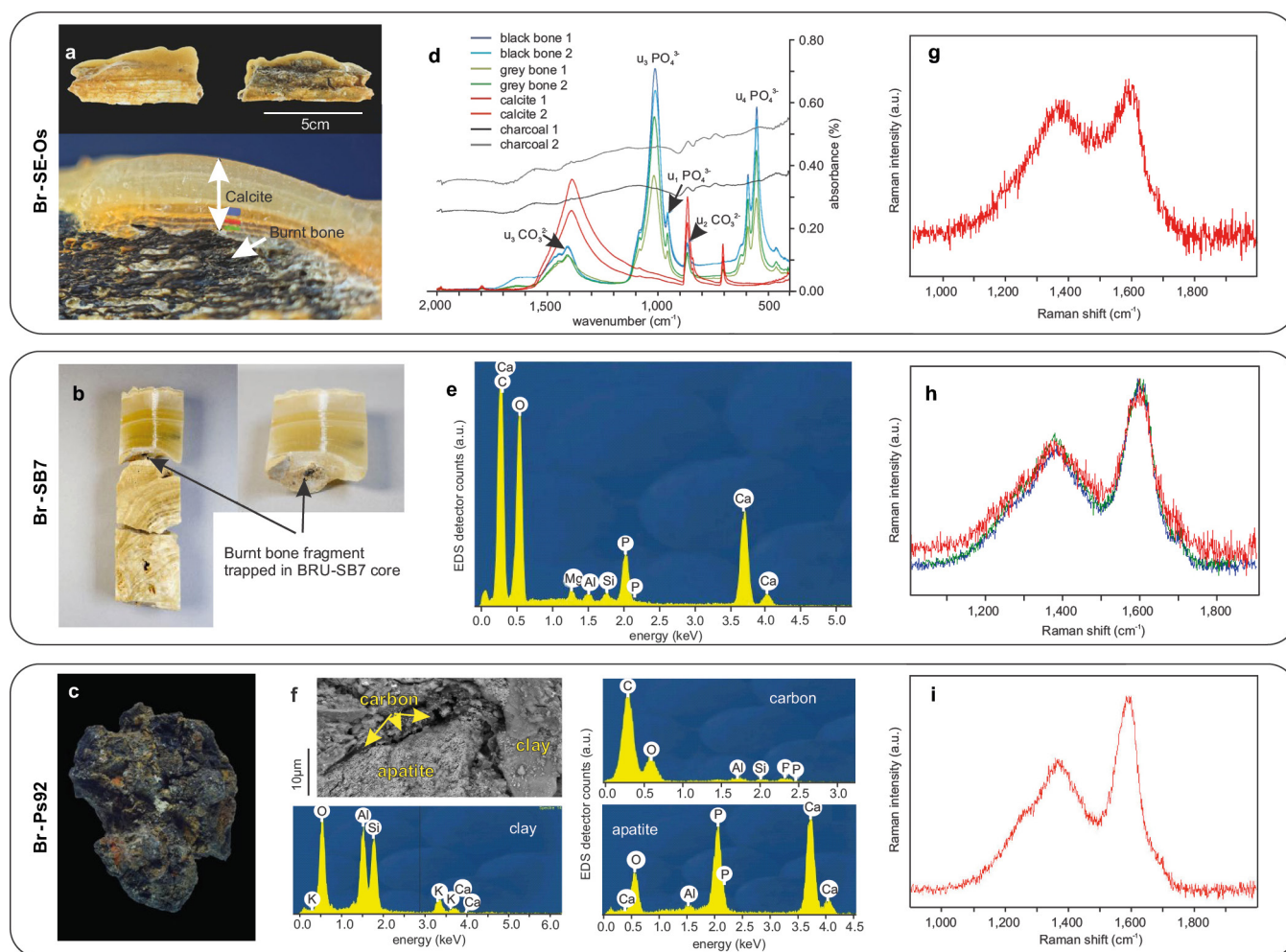
Extended Data Figure 4 | Statistics of the speleofacts. **a, b**, Kernel density estimates for the dimensions (**a**, length and **b**, diameter) of speleofacts across the different structures. Structure A can be distinguished from the others by the presence of very large speleofacts. Such speleofacts are not present in structure D, and only rarely in structure B. Structure C, despite its very small size, is worth considering due to the large dimensions of its speleofacts. Structures E and F, with only a few speleofacts and no specific features, are not represented here. A Kruskal–Wallis test conducted on the structures represented here shows a significant difference between the median length and median diameter across structures ($P < 0.05$). A post hoc analysis of the diameter with Hochberg’s adjustment method, distinguishes structure C from the three others. **c**, The weight of the speleofacts is estimated by the following formula: $\pi D^2 L \rho / 12 \times (1 + d/D + d^2/D^2)$ where D is the maximum diameter, d the minimum diameter, L the maximal length, and ρ the calcite density. These weights can be roughly estimated by considering them as truncated cones. As their maximum

length L , maximum diameter D , and minimum diameter d are known, their volume can be easily estimated (Extended Data Table 1). Their weight is then obtained by multiplying the previous quantity by the calcite density ρ , which is comprised between 2.5 and 2.8 g cm⁻³ depending on its porosity and detrital contamination. Minimal weights are obtained using a density of 2.5 g cm⁻³. **d**, The figure shows the mean weights and their 95% confidence interval in each structure. **e–g**, The orientation data (Schmidt diagram³⁶) of the speleofacts in the three main structures (A, B, D) are very similar (**e**, structure A; **f**, structure B; **g**, structure D) and do not show any preferential direction. The distance to the centre of the circle represents the slope; the distribution of the speleofacts is isotropic and mostly planar. This confirms in all cases that such orientation and slope patterns cannot be due to natural processes related to water flow, mass flows or other gravitational processes³⁷, which in any case would not have resulted in the current geomorphology of the cave in this sector.



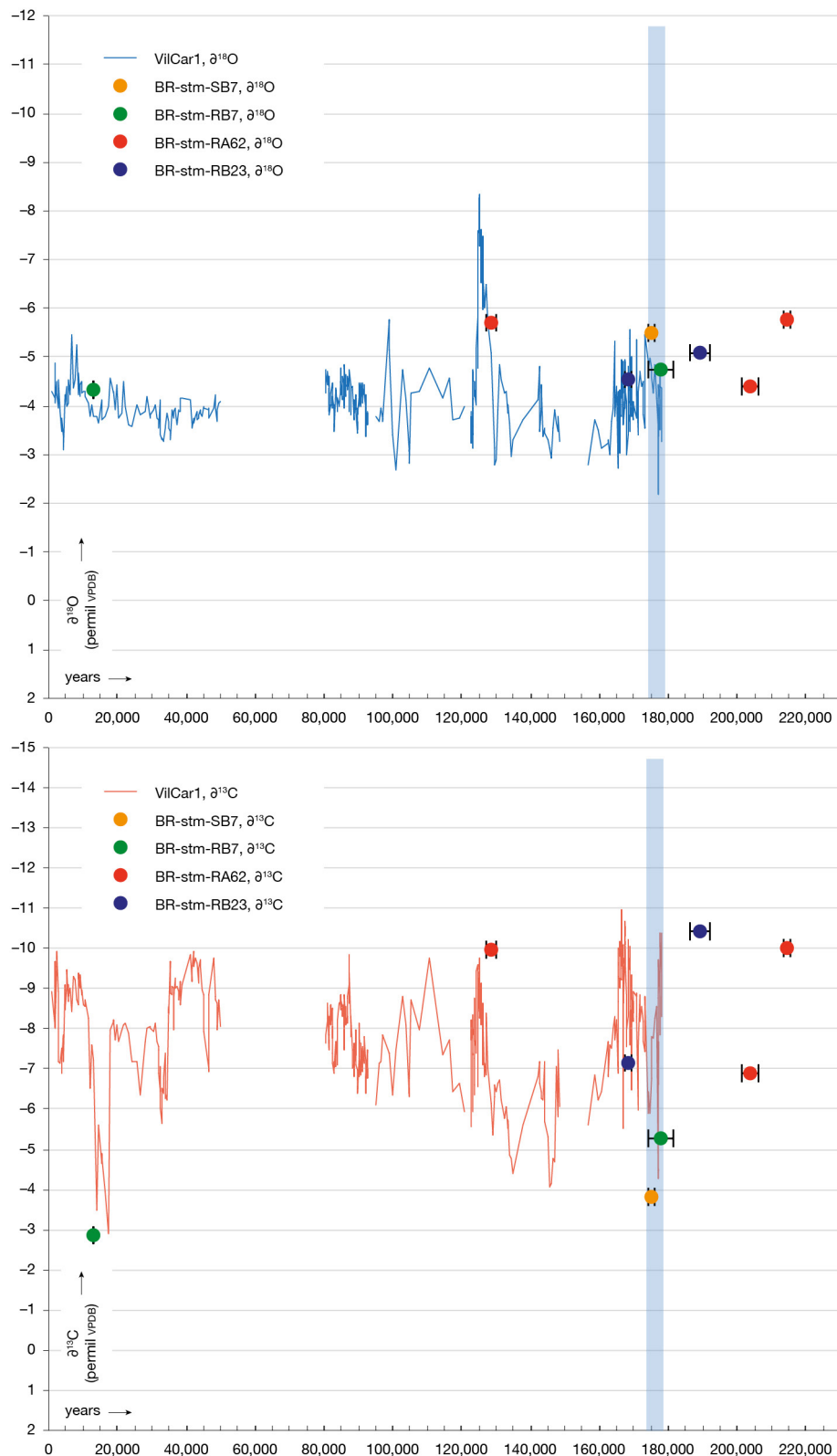
Extended Data Figure 5 | Magnetic survey above the structures. Red circles: main recognized hearths. The magnetic survey aims to reveal the locations that were heated, including hearths or smaller fireplaces through the detection of magnetic anomalies. The first archaeological applications of this prospection method are for the location of heated archaeological structures (see pages 422–519 of ref. 38). The magnetic properties enhancement by heating was first demonstrated for soils^{39–41}, and then on substrate of caves^{42–44}. In this type of hydromorphic environment, iron is present as nonmagnetic or weak magnetic FeOOH minerals, such as goethite (see pages 375–421 of ref. 38). In these conditions, temperature elevation above 200–250 °C induces dehydration of the FeOOH, present in clay material, to Fe₃O₄ (magnetite) which is a strong magnetic mineral⁴³. The increase of magnetic susceptibility induced by heating offers similar information than thermoluminescence methods⁴³. In the present case, a magnetic susceptibility increase beyond a factor of two was observed after heating a clay sample of the cave. Therefore, the heated clay-like material, even if present only in small amounts in speleothems, acquired a sufficiently high magnetization to generate a local earth magnetic deformation, also called an anomaly. As this deformation decreases when the source distance increases (see pages 422–519 of ref. 38), a larger anomaly with a medium intensity might reveal a hearth under the stalagmitic floor (between structures B and C), calcite being magnetically nearly neutral (diamagnetic). The realization of magnetic survey at high spatial resolution for detection of paleohearths in prehistoric cave is a recent innovation⁴⁴. The magnetic field explored above the structures was over one metre thick, with a dual sensor G858 Geometrics magnetometer with an extended cable. A 360° prism was inserted between both sensors, which were superposed at a distance of 0.22 m. These elements were hung at the end of a telescopic boom pole and fixed on a tripod. 3D geolocation measurements were ensured by tracking with a Trimble S8 total station following the 360° prism. This apparatus allows coverage of a volumetric

space up to 5 m from the operator with ten measurements per second while controlling the space covered⁴⁴. Extended Data Fig. 5 presents the results of the magnetic measurements. Altitude contour lines (8.5 cm distance interval) are extracted from photogrammetric data. The magnetic intensity point cloud is a bottom view of the magnetic field intensity gradient, that is, the difference in magnetic field intensity as measured between the bottom and top sensors. As the local past and present magnetic field have an inclination of ~63° down, a magnetic source generates a dipolar local deformation of the magnetic field with a negative anomaly to the north and positive to the south³⁸. In Extended Data Fig. 5, a dipole corresponds to a blue and red spot aligned approximately north–south. The majority of the main dipoles of metric dimension observable are mostly associated to fire traces (reddened, blackened calcite) observed on the horizontally positioned stalagmites, for example, the heated zone of the structures D and E. Increases of magnetic viscosity, known as a fire marker⁴², are observed in such zones. Some places present split positive anomalies, for example, places located on structure D, indicating twin core fires or non-contemporaneous fires. The main measured dipole is located to the west of structure B at the border of a zone covered by a calcite layer and near a char concentration zone, which suggests the occurrence of a hearth underneath the flowstone. Some visible heated zones did not reveal any magnetic anomaly, indicating that the substratum at these places was heated below 200–250 °C. The most tenuous dipoles located on the flat ground surface may reflect the changing nature of the substratum, rather than any heating. Indeed, the weak magnetic contrast between clay material and calcite material can be the source of a weak anomaly. An alternative explanation is the presence of a heated zone underneath a thick stalagmitic floor, the distance between source and measurement mitigating the anomaly³⁸. For example, an anomaly located at midway between structures B and C. Complementary analysis of the spatial distribution of the clay material must be realized to determine which hypothesis is correct.



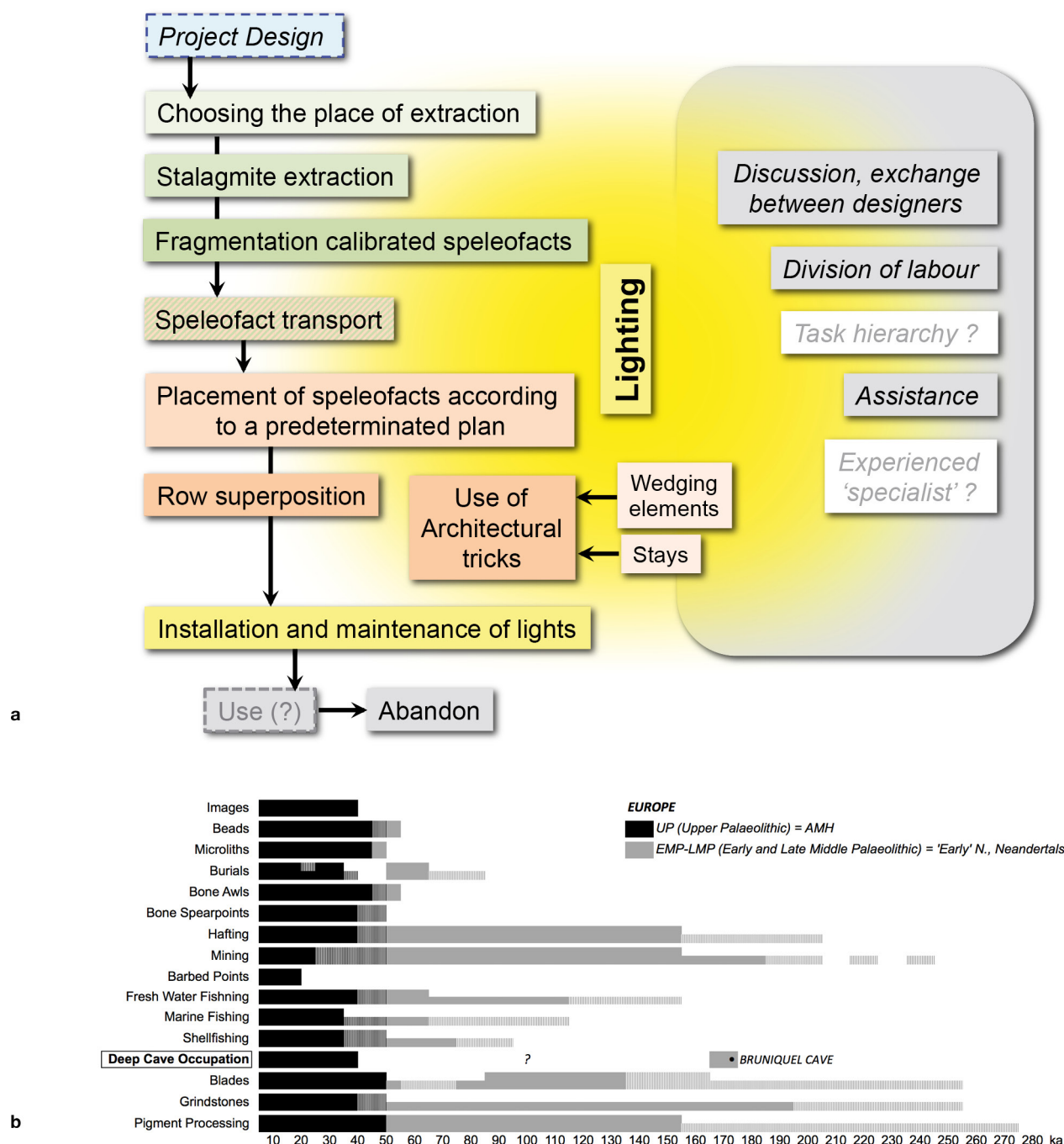
Extended Data Figure 6 | Burnt bone fragments. Three black fragments (a, b, c) were analysed with a scanning electron microscope energy dispersive spectrometry probe (SEM-EDS) (e, f), fast Fourier infra-red, FTIR (d) and Raman spectrometry (g, h, i). FTIR analyses were made at the Laboratoire de Physique des Solides (LPS), Paris-XI University, Orsay by S. Mariot on a Nicolet iS50 ABX spectrometer. Raman spectroscopy was performed with an Invia spectrometer from Renishaw and the atomic spectrometry was performed with a FE-SEM Zeiss Sigma equipped with an EDS probe at the École Normale Supérieure, Paris, France. **a**, A 6.7-cm-long piece of burnt bone (Br-SE-Os) trapped between stalagmite elements in structure E (Extended Data Fig. 5) was almost completely covered by calcite except on its medullar side. Three layers were sampled for uranium-series dating (green, red and blue marks) (Extended Data Table 2). The bone with the 5-mm-thick calcite crust was cut longitudinally and the calcite was sampled along deposition layers, starting at the internal surface after removing the bone material. Three thin discontinuities marked by thin brownish layers separate the deposits into three calcite layers from which three ^{230}Th samples were taken (Extended Data Table 2). Except the middle sub-sample, which was contaminated by detrital elements (high ^{232}Th concentration), ^{230}Th ages given by the other two sub-samples are in stratigraphic order and in agreement with the age of the structures. This demonstrates that humans introduced this bone before 180.9 ± 20.3 ka. Note the elongated medulla cells of the bone and their deep black colour, suggesting that the collagen was carbonized at a temperature between 300 and 400°C ^{45,46}. Note that the burnt bone was covered by a reddish and blackened speleofact (Extended Data Fig. 5), due to the heat. **d**, FTIR spectroscopy (blue spectrum on the black part of the bone, green spectrum on the grey part of the bone, red spectrum on the overlying calcite crust and grey spectrum on a modern char) show well-characterized PO_4^{3-} absorbance peaks, suggesting that the bone was burnt; such as the slightly more individualized peak at $\sim 618\text{ cm}^{-1}$; and the splitting factor (SF)

calculated with the heights of the 603 and 565 cm^{-1} peaks, which are here relatively high (4.6 to 4.8) and typical of burnt bones⁴⁷. **g**, Raman spectrometry displays two well-defined peaks at $1,580\text{ cm}^{-1}$ and at $1,350\text{ cm}^{-1}$, characteristic of char, demonstrating that it was burnt^{48,49}. **b**, Sample Br-SB7 is a 3 mm large black fragment found trapped in the core of Br-stm-SB7 (Fig. 2). This fragment is situated just below the base of the regrowth dated to 175.2 ± 0.8 ka, and just above the ancient surface of the 'old' stalagmite (whose layers have been dated to 222.4 ± 5.8 ka). **h**, Raman spectra of this black fragment display two well-defined peaks at $1,580\text{ cm}^{-1}$ and at $1,350\text{ cm}^{-1}$ characteristic of char carbon^{49,50}. **e**, SEM-EDS shows the presence of phosphorous, in addition to carbon, suggesting that it is a burnt bone fragment, similar to the larger bone piece (a). Because it is trapped in the dated calcite core, it also demonstrates that the fire occurred before 175.2 ± 0.8 ka. **c**, A black aggregate of millimetre-sized fragments (Br-PS92), mainly burnt bones of 1–3 cm was collected in 1992 by F. Rouzaud in the char concentration zone near structure B (Extended Data Fig. 5), and analysed recently. **i**, As with the previous samples, the Raman spectrum is typical of char carbon with vibrational bands at $1,580\text{ cm}^{-1}$ and $1,350\text{ cm}^{-1}$. **f**, The SEM images (back scattered mode) show a blend of at least three phases at the micrometre scale. The elemental analyses performed by EDS on each of these phases allow their attribution to a carbonaceous component (the EDS spectrum shows a major peak of carbon), a phosphorous component (the three major peaks (Ca, P and O) strongly suggest a phase belonging to the apatite family), and a clay component (attested by the coexistence of the three major peaks Si, Al, O), respectively. The Raman spectra demonstrate that the carbonaceous component is a char^{48,49}, that is, a carbonaceous solid resulting from the heat treatment of an organic precursor. These results confirm that the char concentration zone near structure B was most probably a hearth, and that humans burned bones on the clay-like soil of the cave.



Extended Data Figure 7 | Calcite core stable isotope results. a, b, Stable isotope measurements (calcite $\delta^{18}\text{O}$ and $\delta^{13}\text{C}$) were made on parts of cores extracted from the structure to check the coherency of the isotope signal with an already published time series from speleothems from the Villars Cave (Dordogne)⁵⁰, located 100 km to the northwest of Bruniquel Cave. The results reveal a good match between the average $\delta^{18}\text{O}$ of regrowths after 176 ka and the Vil-car1 flowstone stable isotopes. This is also true for the sample that covers marine isotope stage 5e, with a much lower amplitude change, however. The Bruniquel core $\delta^{13}\text{C}$ signal appears more variable, possibly due to a greater sensitivity of the vegetation density to

climatic changes or to detrital contaminations, which are probably close to the discontinuity at the base of the regrowths (b). Higher resolution measurements combined with more uranium-series dating will allow the construction of short palaeoclimatic time series and more detailed observations of climatic variations. Today, the Structure Chamber has an extremely stable temperature of $12.68 \pm 0.02^\circ\text{C}$ (two times the standard deviation of the temperature values measured during one year with a time step of 1.5 h) compared to the outside temperature over the same period ($13.2 \pm 8.8^\circ\text{C}$). These results indicate the current confinement of the cave environment, important for isotopic studies.



Extended Data Figure 8 | Human appropriation of the underground environment: above, the specific task sequence in Bruniquel Cave (a). Below, replacement within the general context of various indicators of modern behaviour (b). **a**, *Chaine opératoire* (task sequence) of the construction of the structures in Bruniquel Cave. This type of construction implies the beginnings of a social organization: this organization could consist of a project that was designed and discussed by one or several individuals, a distribution of the tasks of choosing, collecting and calibrating the speleofacts, followed by their transport (or vice versa) and placement according to a predetermined plan. This work would also require adequate lighting. The construction of such a structure, involving the placement and arrangement of speleofacts, supposes a minimum degree of skill, since architectural techniques such as inserting wedging elements between two rows of speleofacts (Extended Data Fig. 2d, e),

or placing stays to act as buttresses (Extended Data Fig. 2c), appear to have been used. We evaluated the number of speleofacts used (approximately 400), as well as their combined weight (between 2.1 and 2.4 tons), but not yet the number of hours necessary to realize the structures. This will require long and complex experimental procedures that will be undertaken in future research. The complexity of the structure, combined with its difficult access (335 m from the cave entrance), are signs of a collective project and therefore suggest the existence of an organized society that was already on the path to 'modernity'. Until now, no site of this age, attributed to Neanderthals—even late ones—or early modern humans has been associated with such activities in an underground space. **b**, A multiple species model for the origin of behavioural modernity in Europe. Modified from ref. 15, to which was added the 'Deep Cave Occupation' and 'Bruniquel Cave'.

Extended Data Table 1 | Speleofacts: definition and archaeometry

Speleofacts		Structures						Total
		A	B	C	D	E	F	
Number	Number of speleofacts	267	49	9	53	15	6	399
	% / total	66.92%	12.28%	2.26%	13.28%	3.76%	1.50%	100
Length (m)	Total length of the speleofacts (m)	83.74	13.95	3.28	11.17	4.14	1.40	117.68
	% / total	71.16%	11.85%	2.79%	9.50%	3.52%	1.19%	100
	Average length	31.7	28.5	36.4	21.5	27.6	23.3	29.8
Weight (kg)	Weight (kg)	high estimate	1,771.28	280.28	77.97	164.65	63.45	2,385.04
		low estimate	1,581.50	250.27	69.62	147.01	56.65	2,129.52
	% / total		74.27%	11.75%	3.27%	6.90%	2.66%	100
	Average weight		6,251	5,107.5	8,702.4	3,127.8	4,046.3	6120
Diameter Ø (cm)	Ø "root" (extraction)		16.50	15.48	28.50	13.94	16.50	–
	maximum Ø		9.41	8.91	11.17	8.54	7.95	9.33
	minimum Ø		7.17	7.16	9.02	7.14	6.80	8.87

A speleofact is defined as any element extracted from a speleothem (stalagmite, stalactite, drapery, flowstone, stalagmitic column, etc.) with the intent to use it for a precise purpose, thus removing it from its original formation location. This use is linked to a human activity, such as in the realization of any type of modification or construction, use as a utensil or for decoration, or for any other purpose. From the moment it is collected, the element in question attains a status that is distinct from its natural formation context, whether or not it is transformed by flaking, shaping, retouching, striking, engraving, painting, etc. Speleothems that have been clearly worked (flaked, shaped, retouched, pecked, engraved, painted, etc.) while remaining *in situ* should also be included in this definition.

Extended Data Table 2 | Speleothem ^{230}Th dating results

Core Number	Sample Number	^{238}U (ppb)	^{232}Th (ppt)	$^{230}\text{Th} / ^{232}\text{Th}$ (atomic $\times 10^{-9}$)	$\delta^{234}\text{U}^*$ (measured)	$^{230}\text{Th} / ^{238}\text{U}$ (activity)	^{230}Th Age (ky) (uncorrected)	^{230}Th Age (ky) (corrected)	$\delta^{234}\text{U}_{\text{initial}}^{***}$ (corrected)	^{230}Th Age (ky BP)*** (corrected)
Speleothems in the structures ('speleofacts')										
BR-stm-SA249	2491	205.4 \pm 0.3	16,704 \pm 335	256 \pm 5	466.2 \pm 2.4	1.2620 \pm 0.0028	178.5 \pm 1.2	177.1 \pm 1.5	769 \pm 5	177.1 \pm 1.5
BR-stm-RA62	621	90.5 \pm 0.1	14,623 \pm 293	146 \pm 3	898.7 \pm 2.8	1.4299 \pm 0.0032	130.6 \pm 0.6	128.6 \pm 1.6	1,292 \pm 7	128.5 \pm 1.6
BR-stm-RA62	622	154.9 \pm 0.2	30,279 \pm 607	125 \pm 3	593.6 \pm 2.3	1.4766 \pm 0.0027	206.8 \pm 1.3	204.0 \pm 2.3	1,056 \pm 8	203.9 \pm 2.3
BR-stm-RA62	623	133.4 \pm 0.1	4037 \pm 81	811 \pm 16	583.5 \pm 1.4	1.4885 \pm 0.0019	215.0 \pm 0.9	214.6 \pm 1.0	1,069 \pm 4	214.6 \pm 1.0
BR-stm-RA62	624	98.9 \pm 0.1	1131 \pm 23	2,309 \pm 46	573.1 \pm 1.7	1.6008 \pm 0.0019	273.6 \pm 1.7	273.5 \pm 1.7	1,240 \pm 7	273.4 \pm 1.7
BR-stm-SB7	71	330.8 \pm 0.8	150,823 \pm 3,036	50 \pm 1	462.0 \pm 3.3	1.3907 \pm 0.0044	229.8 \pm 2.7	222.5 \pm 5.8	866 \pm 15	222.4 \pm 5.8
BR-stm-SB7	72	134.8 \pm 0.2	242,876 \pm 4,870	13 \pm 1	306.0 \pm 2.5	1.4098 \pm 0.0086	484.4 \pm 40.1	456.8 \pm 38.4	1,111 \pm 122	456.7 \pm 38.4
BR-stm-SB7	73	68.2 \pm 0.1	1,939 \pm 39	809 \pm 16	612.9 \pm 1.7	1.3954 \pm 0.0021	175.7 \pm 0.7	175.3 \pm 0.8	1,005 \pm 4	175.2 \pm 0.8
BR-stm-RB7	74	95.1 \pm 0.1	30,522 \pm 611	70 \pm 1	548.3 \pm 1.8	1.3574 \pm 0.0028	183.0 \pm 1.0	178.0 \pm 3.7	906 \pm 10	177.9 \pm 3.7
BR-stm-RB7	75	246.2 \pm 0.3	31,548 \pm 632	176 \pm 4	401.9 \pm 1.9	1.3684 \pm 0.0019	254.6 \pm 1.8	252.5 \pm 2.3	820 \pm 7	252.5 \pm 2.3
BR-stm-RB7	76	219 \pm 0.2	45,447 \pm 910	106 \pm 2	267.5 \pm 1.6	1.3345 \pm 0.0020	410.1 \pm 7.3	406.8 \pm 7.5	843 \pm 18	406.7 \pm 7.5
BR-stm-RB7	77	133 \pm 0.2	9 \pm 1	48,099 \pm 2,907	801.2 \pm 1.8	0.2040 \pm 0.0012	13.0 \pm 0.1	13.0 \pm 0.1	831	12.9 \pm 0.1
BR-stm-SA59	59	224.8 \pm 0.2	744 \pm 15	8,828 \pm 178	922.1 \pm 1.8	1.7723 \pm 0.0020	193.5 \pm 0.7	193.4 \pm 0.7	1,592 \pm 4	193.4 \pm 0.7
BR-stm-RB23	231	92.5 \pm 0.1	1,595 \pm 32	1241 \pm 25	401.6 \pm 1.7	1.2991 \pm 0.0017	217.9 \pm 1.2	217.7 \pm 1.2	742 \pm 4	217.6 \pm 1.2
BR-stm-RB23	232	87 \pm 0.1	4,889 \pm 98	421 \pm 8	688.1 \pm 2.1	1.4422 \pm 0.0026	169.1 \pm 0.8	168.3 \pm 1.0	1,106 \pm 5	168.3 \pm 1.0
BR-stm-RB23	233	147 \pm 0.2	34,055 \pm 683	91 \pm 2	433.6 \pm 1.7	1.2724 \pm 0.0022	193.1 \pm 1.0	189.2 \pm 3.0	740 \pm 7	189.1 \pm 3.0
BR-stm-RB23	234	76 \pm 0.1	157 \pm 3	10,295 \pm 209	409.7 \pm 1.6	1.2859 \pm 0.0019	208.2 \pm 1.1	208.2 \pm 1.1	737 \pm 4	208.1 \pm 1.1
Flowstone inside Structure A										
BR-PL-P13	13	163.2 \pm 0.2	82,660 \pm 1,655	45 \pm 1	564.0 \pm 1.8	1.3757 \pm 0.0034	183.8 \pm 1.2	175.9 \pm 5.7	927 \pm 15	175.9 \pm 5.7
Stalagmites on the collapsed rocks in the entrance zone										
BR-stm-3	31	41.8 \pm 0.1	2,780 \pm 56	226 \pm 5	203.2 \pm 1.5	0.9104 \pm 0.0020	144.5 \pm 0.8	143.0 \pm 1.3	304 \pm 3	142.9 \pm 1.3
BR-stm-3	32	15.7 \pm 0.1	14,278 \pm 286	17 \pm 1	139.1 \pm 3.7	0.9476 \pm 0.0080	181.2 \pm 4.1	157.1 \pm 17.9	217 \pm 12	157.0 \pm 17.9
BR-stm-2	21	32.5 \pm 0.1	21,647 \pm 434	25 \pm 1	131.7 \pm 2.8	0.9959 \pm 0.0050	211.0 \pm 3.6	194.0 \pm 12.6	228 \pm 9	193.9 \pm 12.6
BR-stm-2	22	28.1 \pm 0.1	13,737 \pm 275	34 \pm 1	143.0 \pm 2.6	0.9975 \pm 0.0035	204.9 \pm 2.5	192.9 \pm 8.9	246 \pm 8	192.8 \pm 8.9
Flowstone on the collapsed rocks at the beginning of the main gallery										
BR-PL-1	1	68.5 \pm 0.1	4,066 \pm 81	316 \pm 6	132.9 \pm 1.4	1.1389 \pm 0.0017	367.4 \pm 5.8	366.1 \pm 5.8	373 \pm 7	366.0 \pm 5.8
Calcite on the burnt bone										
Calos	Calos-1	136 \pm 0.1	210,059 \pm 4,206	15 \pm 0	497.7 \pm 1.9	1.3792 \pm 0.0045	208.2 \pm 1.9	180.9 \pm 20.3	829 \pm 48	180.9 \pm 20.3
Calos	Calos-2	117 \pm 0.2	500,389 \pm 10,028	6 \pm 0	467.3 \pm 2.0	1.5061 \pm 0.0102	296.3 \pm 8.4	198.6 \pm 9.5 $\times 10^{21}$	818 \pm 504	198.5 \pm 9.5 $\times 10^{21}$
Calos	Calos-3	50 \pm 0.1	55,973 \pm 1,122	16 \pm 0	463.7 \pm 2.4	1.0834 \pm 0.0044	132.3 \pm 1.1	110.6 \pm 15.8	634 \pm 28	110.6 \pm 15.8

The table shows the dating of stalagmites used to build the structures (speleofacts), those of the stalagmites that grew on the structures (regrowths, in darker lines), the flowstone inside the main structure A, and the calcite on the entrance collapse. The dating results of the calcite, deposited on the burnt bone found in the structures, are shown in the last lines. One date was rejected (Calos-2) due to its high uncertainty. Corrected ^{230}Th ages assume the initial $^{230}\text{Th}/^{232}\text{Th}$ atomic ratio of $4.4 \pm 2.2 \times 10^{-6}$, which is the value for a material at secular equilibrium, with the bulk earth $^{232}\text{Th}/^{238}\text{U}$ value of 3.8. Errors are arbitrarily assumed to be 50%. Age uncertainties are given as 2σ .

p.p.b., parts per billion, 1×10^{-9} ; p.p.t., parts per trillion, 1×10^{-12} ; BP, before present, with the present defined as 1950 AD.

A shared neural ensemble links distinct contextual memories encoded close in time

Denise J. Cai^{1*}, Daniel Aharoni^{1,2,3*}, Tristan Shuman^{2,3*}, Justin Shobe^{1*}, Jeremy Biane^{4,5}, Weilin Song¹, Brandon Wei¹, Michael Veshkini¹, Mimi La-Vu¹, Jerry Lou^{2,3}, Sergio E. Flores^{2,3}, Isaac Kim¹, Yoshitake Sano¹, Miou Zhou¹, Karsten Baumgaertel⁶, Ayal Lavi¹, Masakazu Kamata⁷, Mark Tuszyński^{4,5}, Mark Mayford⁶, Peyman Golshani^{2,3} & Alcino J. Silva¹

Recent studies suggest that a shared neural ensemble may link distinct memories encoded close in time^{1–12}. According to the memory allocation hypothesis^{1,2}, learning triggers a temporary increase in neuronal excitability^{13–15} that biases the representation of a subsequent memory to the neuronal ensemble encoding the first memory, such that recall of one memory increases the likelihood of recalling the other memory. Here we show in mice that the overlap between the hippocampal CA1 ensembles activated by two distinct contexts acquired within a day is higher than when they are separated by a week. Several findings indicate that this overlap of neuronal ensembles links two contextual memories. First, fear paired with one context is transferred to a neutral context when the two contexts are acquired within a day but not across a week. Second, the first memory strengthens the second memory within a day but not across a week. Older mice, known to have lower CA1 excitability^{15,16}, do not show the overlap between ensembles, the transfer of fear between contexts, or the strengthening of the second memory. Finally, in aged mice, increasing cellular excitability and activating a common ensemble of CA1 neurons during two distinct context exposures rescued the deficit in linking memories. Taken together, these findings demonstrate that contextual memories encoded close in time are linked by directing storage into overlapping ensembles. Alteration of these processes by ageing could affect the temporal structure of memories, thus impairing efficient recall of related information.

Contextual memories are encoded in discrete and sparse populations of neurons in the hippocampus^{17–21}. Recent findings demonstrated that increasing the relative neuronal excitability of a subset of neurons increases the probability that those neurons will participate in a memory trace^{6,8–11}. While previous studies used viral vectors to manipulate excitability, temporary increases in excitability occur naturally following learning, including in the hippocampus^{13,14,22}. Therefore, two distinct memories could be linked across time because the temporary increase in excitability would bias the storage of a subsequent memory to many of the same neurons that encoded the first memory, such that recall of one of these events would also probably lead to recall of the other, a key prediction of the memory allocation hypothesis^{1,2}.

To investigate the neuronal ensembles encoding multiple memories, we constructed an open-source, head-mounted, miniature fluorescent microscope^{7,23}, to image *in vivo* calcium transients in CA1 neurons using GCaMP6f. With this approach we tracked the activation of the same neurons in mice as they freely explored three distinct novel contexts across multiple days (Fig. 1a–c, Extended Data Figs 1 and 2). We recorded CA1 neurons activated by three different contexts separated

by either 5 h or 7 days. Previous studies show transient learning-dependent increases in neuronal excitability^{13,14,24} and we confirmed that 5 h after context exposure there was an increase in excitability in CA1 neurons that encoded the context (Extended Data Fig. 3c, d). Therefore, we predicted that the overlap between the neural representations of two contexts separated by 5 h would be higher than the overlap of the neural representations of two contexts separated by 7 days.

We exposed mice to three distinct, novel contexts. A and C were separated by 7 days; B and C were separated by 5 h. Using miniature microscopes, we imaged active CA1 neurons during each context exploration (Fig. 1d). We found more overlap between the neural ensembles encoding B and C, spaced 5 h apart, than between the neural ensembles encoding A and C, spaced 7 days apart (Fig. 1f, Extended Data Fig. 4a, b). Notably, this difference was not due to differences in the total number of active CA1 cells in the three contexts (Fig. 1e). We confirmed these findings with the TetTag transgenic system, a non-invasive technique that allowed us to tag neurons active during the exploration of two contexts^{25,26} (Fig. 2a, b, Extended Data Fig. 3a, b). We used this transgenic approach to tag the neural ensemble activated by exploration of an initial novel context (GFP⁺) and compared this population to the ensemble activated by exploration of a second distinct, novel context (using ZIF immunohistochemistry), either 5 h or 7 days later (Fig. 2c–e). When the two contexts were separated by 7 days, the overlap between the two ensembles was similar to what was expected due to chance (Fig. 2f), indicating that independent populations of neurons encoded the two distinct contexts. However, when the two contexts were separated by 5 h, overlap between neuronal ensembles was significantly above chance levels and higher than in the 7 days group (Fig. 2f). Together, the calcium imaging and TetTag data provide converging evidence that overlapping neural ensembles encode distinct contexts when these contexts are separated by 5 h, but not by 7 days.

To determine whether the overlap of neuronal representations link contextual memories that occurred close in time, such that the recall of one is more likely to lead to the recall of the other, we again exposed animals to three distinct contexts as described above: A and C were separated by 7 days, and B and C were separated by 5 h. Two days later, mice were placed in C and given an immediate footshock (Fig. 3a). Since the neural representations of B and C overlap more than A and C (Extended Data Fig. 5), recall of C (shocked context) should lead to recall of B (but not A). Therefore, the fear associated with C should transfer to B (but not to A). Remarkably, we found that mice tested in B, a context in which they had not been shocked, froze as much as mice tested in C (shocked context; Fig. 3b). In contrast, mice tested in A froze significantly less than mice tested in the other two contexts.

¹Departments of Neurobiology, Psychiatry & Biobehavioral Sciences and Psychology, Integrative Center for Learning and Memory, Brain Research Institute, University of California, Los Angeles, California 90095, USA. ²Departments of Neurology and Psychiatry & Biobehavioral Sciences, Integrative Center for Learning and Memory, Brain Research Institute, University of California, Los Angeles, California 90095, USA. ³West Los Angeles VA Medical Center, 11301 Wilshire Blvd, Los Angeles, California 90073, USA. ⁴Department of Neurosciences, University of California, San Diego, La Jolla, California 92093, USA. ⁵Veterans Affairs Medical Center, San Diego, California 92161, USA. ⁶Departments of Cell Biology and Neurosciences, Institute for Childhood and Neglected Diseases, The Scripps Research Institute, 10550 North Torrey Pines Road, La Jolla, California 92037, USA. ⁷Division of Hematology/Oncology, David Geffen School of Medicine, University of California, Los Angeles, California 90095, USA.

*These authors contributed equally to this work.

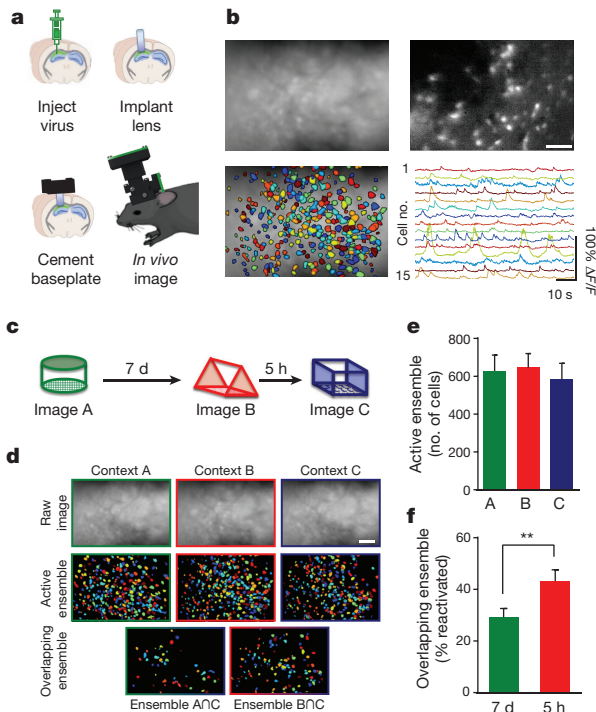


Figure 1 | Calcium imaging CA1 with integrated miniature microscopes while exploring different contexts. **a**, A microendoscope was implanted directly above CA1 expressing viral GCaMP6f and a baseplate was affixed onto the skull. A miniature fluorescent wide-field microscope was used to image CA1 neurons across repeated imaging sessions. **b**, Top left, example image of mean fluorescence during context exploration. Top right, example image of relative fluorescent change ($\Delta F/F$). Bottom left, cells extracted from $\Delta F/F$. Scale bar represents 100 μm . Bottom right, example traces of $\Delta F/F$ colour coded to represent individual neurons. **c**, Experimental design. Mice were imaged while exploring three novel contexts (A, B, C) separated by 7 days or 5 h. **d**, Representative imaging during context exploration. Top row, images of mean fluorescence from each session. Middle row, ensemble of cells active in each session. Bottom row, cells that were active in two sessions. Scale bar represents 100 μm . **e**, There was no difference in the number of cells active across the three context explorations (one-way, repeated measures ANOVA, $F_{2,7} = 2.14$, not significant, $n = 8$ mice). **f**, There was an increase in the overlapping ensemble when contexts are separated by 5 h compared to 7 days (paired t -test, $t_7 = 3.830$, $P = 0.0065$, $n = 8$). $**P < 0.01$. Results show mean \pm s.e.m.

These results support the hypothesis that the overlap between neuronal representations contextually links memories close in time.

Next, we tested whether the memories for B and C remain distinct, rather than forming a unitary memory. If so, extinction of the fear associated with B should not affect recall in C. Again, we exposed animals to B and 5 h later to C, and then two days later paired C with a footshock. Two days after the footshock, the mice were tested in either C (shocked context), B (5 h; not shocked), or D (novel context; Fig. 3c). Consistent with the prior experiment, mice froze similarly in C and B, despite never having been shocked in B. However, they froze less in a novel context (D; Fig. 3d, Extended Data Fig. 6b), demonstrating memory specificity. Next, we carried out repeated exposures in either context C, B, or D daily for 5 days. On the final day, the mice were tested in C (shocked context). As expected, repeated exposures in C (compared to repeated exposures in novel context D) resulted in lower freezing during the extinction test (Fig. 3e). Mice that were repeatedly exposed to B did not show less freezing in C, demonstrating that repeated exposures in B do not cause extinction in C. These results demonstrate that although the memories for B and C show considerable overlap in their ensembles, and recall of B appears to trigger recall of C, memories for these two contexts, acquired 5 h apart, remain distinct.

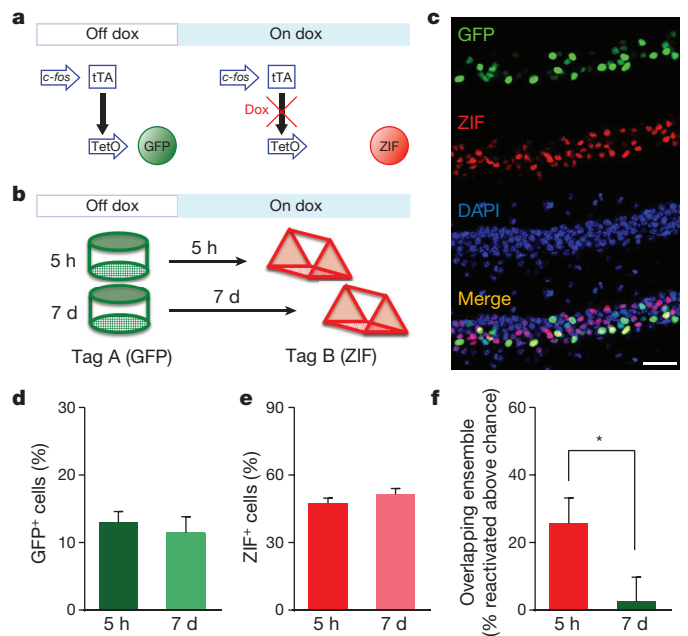


Figure 2 | Tagging neural ensembles of contextual memories with the TetTag system. **a**, Schematic design of the TetTag system. Dox, doxycycline; tTA, tetracycline-transactivator; TetO, tetracycline response element. **b**, Experimental design. Cells active in context A were tagged with GFP and cells active in context B, either 5 h or 7 days later, were labelled with ZIF immunohistochemistry. **c**, Representative examples of GFP, ZIF, DAPI and merged images of CA1. Scale bar represents 50 μm . **d**, There was no difference between the percentage of cells positive for GFP (unpaired t -test, $t_{24} = 0.54$, not significant, $n = 15$, 11 mice). **e**, There was no difference between the percentage of cells positive for ZIF (unpaired t -test, $t_{24} = 1.11$, not significant, $n = 15$, 11 mice). **f**, There was an increase in the overlapping ensemble between contexts when spaced 5 h apart compared to 7 days apart (unpaired t -test, $t_{24} = 2.15$, $P = 0.0422$, $n = 15$, 11 mice). The level of the overlapping ensemble for the 5 h group was above chance (one-sample t -test against 0, $t_{14} = 3.402$, $P = 0.0043$) and at chance for the 7 day group (one-sample t -test against 0, $t_{10} = 0.323$, not significant). $*P < 0.05$. Results show mean \pm s.e.m.

Recent findings demonstrated that manipulations that enhance neuronal excitability can lead to increases in memory strength¹¹. We found that 5 h after exposure to a context, there was an increase in excitability in cells that encoded that context (Extended Data Fig. 3c, d). Thus, the sharing of the neural ensemble and the increase in excitability should result in the strengthening of the memory for a second context 5 h later. To test for modulation of memory strength, mice were exposed to B and then exposed to C 5 h or 7 days later. Two days later, animals received an immediate shock in C. Two days after that, they were tested in C. Home cage controls were trained in the same manner, except they were not exposed to B (Fig. 3f). Mice trained with the 5 h interval had enhanced memory for C compared to either mice trained with the 7 day interval or home cage controls (Fig. 3g; Extended Data Figs 6c, d and 7). Furthermore, this enhancement required NMDA-receptor activity (Extended Data Fig. 8). These data support our previous findings and indicate that for a period of time (5 h, but not 7 days) the processes triggered by the encoding of one memory can modulate the strength of subsequent memories.

Taken together, the results presented above demonstrate that the overlap between the neuronal ensembles representing two separate contextual memories leads to linking of these memories and suggests that excitability has a key role in this process. Since CA1 neuronal excitability decreases with ageing^{15,16,22,27}, we predicted that memory-linking processes may be disrupted in older mice. To test this, we started by repeating the calcium imaging (Fig. 4a) as well as the TetTag experiment (Extended Data Fig. 9e, f) in aged mice. Unlike in young adult mice (3–6 months old), in aged mice (14–18 months old) there

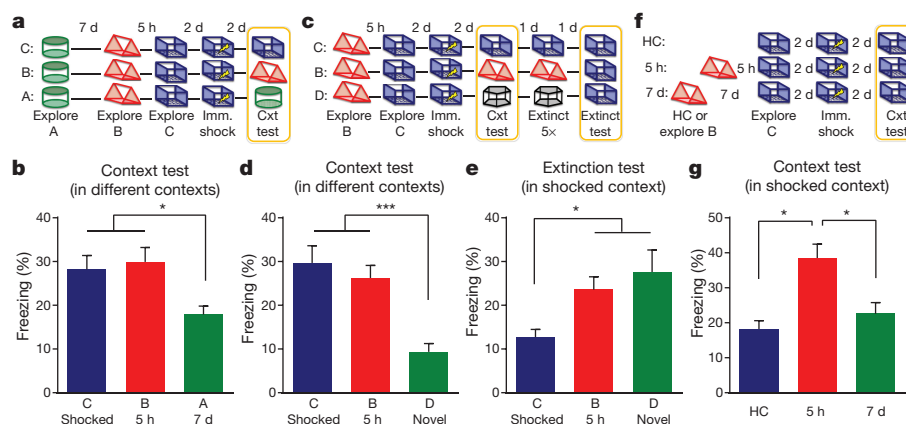


Figure 3 | Memories are contextually linked but distinct. **a**, Design for transfer of fear experiment. Imm. shock, immediate shock; cxt test, context test. **b**, There was a significant difference in freezing between groups that were tested in different contexts (A, B, C) for the transfer of fear experiment (one-way ANOVA, $F_{2,47} = 4.62$, $P = 0.01$, $n = 18, 17, 15$ mice). There was no difference between freezing in contexts C and B ($t_{45} = 0.42$, not significant). Animals had less freezing in context A than C ($t_{47} = 2.46$, $P = 0.02$) and B ($t_{47} = 2.83$, $P = 0.007$). **c**, Design for extinction experiment. **d**, There was a significant difference in freezing during the context test (one-way ANOVA, $F_{2,57} = 12.99$, $P < 0.0001$, $n = 20, 20, 20$ mice). There was no difference between freezing in contexts C and B ($t_{57} = 0.80$, not significant). Animals had less freezing in context D than C ($t_{57} = 4.76$,

$P < 0.0001$) and B ($t_{57} = 3.96$, $P = 0.0002$). **e**, There was a significant difference in freezing during the extinction test (one-way ANOVA, $F_{2,57} = 4.79$, $P = 0.01$, $n = 20, 20, 20$ mice). There were no differences in freezing between groups B and D ($t_{57} = 0.81$, not significant). Group C had less freezing than groups B ($t_{57} = 2.18$, $P = 0.03$) and D ($t_{57} = 2.99$, $P = 0.004$). **f**, Design for enhancement experiment. **g**, There was a significant difference in freezing in the enhancement experiment (one-way ANOVA, $F_{2,51} = 9.63$, $P < 0.001$, $n = 14, 20, 20$ mice). The 5 h group had more freezing than the home cage (HC) ($t_{51} = 3.98$, $P = 0.0002$) and 7 day ($t_{51} = 3.45$, $P = 0.001$) groups. There was no difference between home cage or 7 d groups ($t_{51} = 0.86$, not significant). * $P < 0.05$, ** $P < 0.01$. Results show mean \pm s.e.m.

was no difference between the overlap of neural ensembles encoding contexts spaced 5 h or 7 days apart (Fig. 4b). This lack of overlap was not due to an inability to reliably reactivate the same neural ensemble during recall of the same context (Extended Data Fig. 9a, b) or to general contextual memory deficits (Extended Data Fig. 9c, d).

The results presented above predict that the lack of a shared neural representation in aged mice should disrupt memory linking. To test this hypothesis, we repeated in aged mice the experiment testing the transfer of fear between contexts (Fig. 4c). The results showed that the fear associated with C does not transfer to B in aged mice: the freezing triggered by B (no shock context) was not different than that observed in a novel context, D, and significantly lower than that in C (shocked context; Fig. 4d). Similarly, we found that, unlike in young mice, in aged mice exposure to B (5 h before exposure to C) does not enhance memory for C (Fig. 4e, f). Importantly, this was not due to a deficit in

learning of a single context, since when trained with a single context the performance of aged mice was indistinguishable from that of young mice (Extended Data Fig. 9c, d). Furthermore, the differences between young and aged mice were also not due to strain differences, as we replicated the transfer and enhancement experiments with young mice from the same genetic background as the aged mice (Extended Data Fig. 6). Altogether, these results strongly support the role of neuronal excitability in linking distinct contextual memories encoded close in time, as aged mice exposed to two contexts close in time did not show the increased overlap between ensembles which presumably led to the lack of both the transfer of fear between contexts and the strengthening of the second memory.

To increase neuronal excitability and rescue the memory-linking deficit in aged mice, we injected a lentivirus to express hM3Dq designer receptors exclusively activated by designer drugs (DREADD) tagged

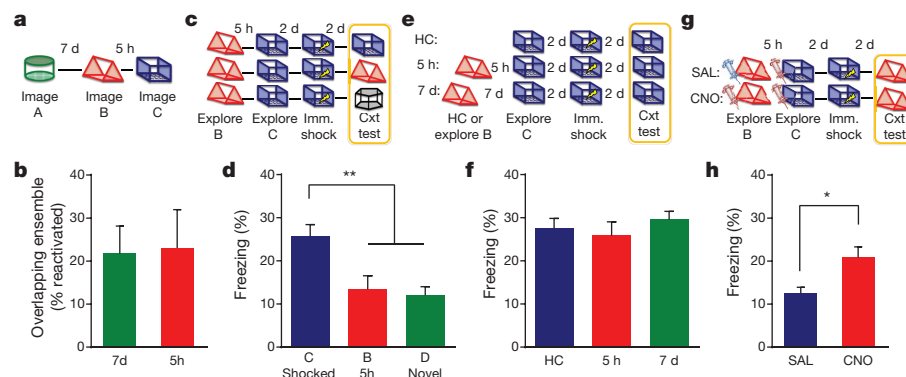


Figure 4 | Age-related deficits in memory linking are rescued by ensemble activation. **a**, Design for calcium imaging with miniature microscope in aged mice. **b**, There was no difference in the overlapping ensemble between the 5 h and 7 day groups (paired t -test, $t_3 = 0.367$, not significant, $n = 4$). **c**, Design for transfer of fear experiment. **d**, There was a significant difference in freezing during the context test (one-way ANOVA, $F_{2,47} = 8.083$, $P = 0.001$, $n = 19, 15, 16$ mice). There was no difference between freezing levels in contexts B and D ($t_{47} = 0.35$, not significant). Animals had more freezing in context C than B ($t_4 = 3.19$,

$P = 0.0025$) and D ($t_{47} = 3.619$, $P = 0.0007$). **e**, Design for behavioural enhancement experiment. **f**, There was no difference in freezing between groups (one-way ANOVA, $F_{2,39} = 0.453$, not significant, $n = 15, 15, 12$ mice). **g**, Design for memory linking rescue by activating cells with DREADD receptors. **h**, There was higher freezing in the CNO group compared to the saline-injected (SAL) group (unpaired t -test, $t_{31} = 2.36$, $P = 0.02$, $n = 12, 21$ mice). * $P < 0.05$, ** $P < 0.01$. Results show mean \pm s.e.m.

with GFP in a sparse population of dorsal CA1 neurons (Extended Data Fig. 10a, b). Clozapine-*N*-oxide (CNO) increases excitability and activates cells that express the DREADD receptors¹¹ (Extended Data Fig. 10c, d). To bias the allocation of the two contextual memories so that they shared an overlapping neural ensemble, we injected CNO before both learning experiences, spaced 5 h apart (Fig. 4g). The control group was given a saline injection before the first exploration and a CNO injection before the second exploration. To test the behavioural consequences of sharing a neural ensemble, mice were brought back two days later for an immediate shock in the second context. Two days later, mice were tested in the first (non-shocked) context to assess their transfer of fear. The CNO group froze more than the saline-injected group in the non-shocked context (Fig. 4h). This was not due to increased anxiety caused by CNO (Extended Data Fig. 10e, f). Thus, increasing neuronal excitability in aged mice rescued the memory-linking deficit.

Mechanisms that link memories are critically important for organizing the enormous number of related memories stored throughout a lifetime. Our results support the memory allocation hypothesis^{1,2} and are consistent with human data and computational modelling²⁸, suggesting that memories encoded within close temporal proximity are more likely to be co-recalled than memories encoded across more distant time frames. Our data indicate that overlapping populations of CA1 neurons serve to link and strengthen memories, thus facilitating integrated recall of experiences encoded close in time while separating those encoded further in time. Temporary increases in excitability^{13–15} probably represent one of a family of mechanisms (synaptic tagging and capture^{2,29} is another example) that structure the acquisition and storage of information to facilitate future use and recall. Alteration of these processes, such as decreases in neuronal excitability during ageing, could affect the organization of memory thus impairing efficient recall of related information.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 12 January; accepted 24 March 2016.

Published online 23 May 2016.

1. Silva, A. J., Zhou, Y., Rogerson, T., Shobe, J. & Balaji, J. Molecular and cellular approaches to memory allocation in neural circuits. *Science* **326**, 391–395 (2009).
2. Rogerson, T. *et al.* Synaptic tagging during memory allocation. *Nat. Rev. Neurosci.* **15**, 157–169 (2014).
3. Mankin, E. A. *et al.* Neuronal code for extended time in the hippocampus. *Proc. Natl Acad. Sci. USA* **109**, 19462–19467 (2012).
4. Han, J.-H. *et al.* Neuronal competition and selection during memory formation. *Science* **316**, 457–460 (2007).
5. Han, J.-H. *et al.* Selective erasure of a fear memory. *Science* **323**, 1492–1496 (2009).
6. Zhou, Y. *et al.* CREB regulates excitability and the allocation of memory to subsets of neurons in the amygdala. *Nat. Neurosci.* **12**, 1438–1443 (2009).
7. Ziv, Y. *et al.* Long-term dynamics of CA1 hippocampal place codes. *Nat. Neurosci.* **16**, 264–266 (2013).
8. Epsztein, J., Brecht, M. & Lee, A. K. Intracellular determinants of hippocampal CA1 place and silent cell activity in a novel environment. *Neuron* **70**, 109–120 (2011).
9. Lee, D., Lin, B. J. & Lee, A. K. Hippocampal place fields emerge upon single-cell manipulation of excitability during behavior. *Science* **337**, 849–853 (2012).
10. Dragoi, G. & Tonegawa, S. Preplay of future place cell sequences by hippocampal cellular assemblies. *Nature* **469**, 397–401 (2011).
11. Yiu, A. P. *et al.* Neurons are recruited to a memory trace based on relative neuronal excitability immediately before training. *Neuron* **83**, 722–735 (2014).
12. Ezzyat, Y. & Davachi, L. What constitutes an episode in episodic memory? *Psychol. Sci.* **22**, 243–252 (2011).

13. Moyer, J. R., Jr, Thompson, L. T. & Disterhoft, J. F. Trace eyeblink conditioning increases CA1 excitability in a transient and learning-specific manner. *J. Neurosci.* **16**, 5536–5546 (1996).
14. McKay, B. M., Matthews, E. A., Oliveira, F. A. & Disterhoft, J. F. Intrinsic neuronal excitability is reversibly altered by a single experience in fear conditioning. *J. Neurophysiol.* **102**, 2763–2770 (2009).
15. Oh, M. M., Oliveira, F. A. & Disterhoft, J. F. Learning and aging related changes in intrinsic neuronal excitability. *Front. Aging Neurosci.* **2**, 2 (2010).
16. Kaczorowski, C. C. & Disterhoft, J. F. Memory deficits are associated with impaired ability to modulate neuronal excitability in middle-aged mice. *Learn. Mem.* **16**, 362–366 (2009).
17. Garner, A. R. *et al.* Generation of a synthetic memory trace. *Science* **335**, 1513–1516 (2012).
18. Guzowski, J. F., McNaughton, B. L., Barnes, C. A. & Worley, P. F. Environment-specific expression of the immediate-early gene *Arc* in hippocampal neuronal ensembles. *Nat. Neurosci.* **2**, 1120–1124 (1999).
19. Liu, X. *et al.* Optogenetic stimulation of a hippocampal engram activates fear memory recall. *Nature* **484**, 381–385 (2012).
20. McKenzie, S. *et al.* Hippocampal representation of related and opposing memories develop within distinct, hierarchically organized neural schemas. *Neuron* **83**, 202–215 (2014).
21. Deng, W., Mayford, M. & Gage, F. H. Selection of distinct populations of dentate granule cells in response to inputs as a mechanism for pattern separation in mice. *eLife* **2**, e00312 (2013).
22. Disterhoft, J. F. & Oh, M. M. Alterations in intrinsic neuronal excitability during normal aging. *Aging Cell* **6**, 327–336 (2007).
23. Ghosh, K. K. *et al.* Miniaturized integration of a fluorescence microscope. *Nat. Methods* **8**, 871–878 (2011).
24. Disterhoft, J. F. & Oh, M. M. Learning, aging and intrinsic neuronal plasticity. *Trends Neurosci.* **29**, 587–599 (2006).
25. Reijmers, L. G., Perkins, B. L., Matsuo, N. & Mayford, M. Localization of a stable neural correlate of associative memory. *Science* **317**, 1230–1233 (2007).
26. Tayler, K. K., Tanaka, K. Z., Reijmers, L. G. & Wiltgen, B. J. Reactivation of neural ensembles during the retrieval of recent and remote memory. *Curr. Biol.* **23**, 99–106 (2013).
27. Murphy, G. G., Shah, V., Hell, J. W. & Silva, A. J. Investigation of age-related cognitive decline using mice as a model system: neurophysiological correlates. *Am. J. Geriatr. Psychiatry* **14**, 1012–1021 (2006).
28. Sederberg, P. B., Howard, M. W. & Kahana, M. J. A context-based theory of recency and contiguity in free recall. *Psychol. Rev.* **115**, 893–912 (2008).
29. Redondo, R. L. & Morris, R. G. Making memories last: the synaptic tagging and capture hypothesis. *Nat. Rev. Neurosci.* **12**, 17–30 (2011).

Acknowledgements We thank B. Khakh for support in the development of the miniaturized microscopes. We thank E. Thai, D. Tarzi, A. Ahuja, K. Lew, E. Lu, E. Stuart, S. Zhang, S. Ghiaee, C. Yang, A. Fariborzi, K. Cheng, N. Rao, A. Chang, C. Grimmick and M. Einstein for help with experiments; N. Rao for assistance with graphical design; and all members of the Silva laboratory for their support. This work was supported by National Institute on Aging R37 AG013622 and the Dr. Miriam and Sheldon G. Adelson Medical Research Foundation to A.J.S.; National Institutes of Health R01 MH101198, 1U54 HD087101 and VA Merit Award BX00152401A1 to P.G.; National Research Service Award F32 MH97413 and Behavioral Neuroscience Training Grant T32 MH15795 to D.J.C.; Neurobehavioral Genetics Training Grant T32 NS048004 and Neural Microcircuits Training Grant T32 NS058280 to D.A.; Cellular Neurobiology Training Grant T32 NS710133 and Epilepsy Foundation Postdoctoral Research Training Fellowship to T.S.; National Institutes of Health U01 NS094286-01 and David Geffen School of Medicine Dean's Fund for development of open-source miniaturized microscopes to A.J.S. and P.G.

Author Contributions D.J.C., J.S., T.S., D.A. and A.J.S. contributed to the study design. D.A., T.S., D.J.C., P.G. and A.J.S. developed the miniature microscope system. D.A. engineered hardware and software associated with the miniature microscope and wrote the MATLAB analysis suite. T.S., D.J.C., W.S., J.S., S.E.F., J.L. and I.K. performed surgeries. D.J.C., T.S., M.L., W.S. and B.W. conducted calcium imaging and TetTag experiments. M.M. engineered and provided TetTag mice. D.J.C., J.S., T.S., M.L., W.S., B.W., M.V. and M.Z. conducted behavioural experiments. D.J.C., D.A., T.S. and A.L. analysed the data. J.B., D.J.C. and T.S. conducted *in vitro* physiology experiments. M.T. supported physiology experiments. Y.S. and M.K. made DREADD virus. D.J.C., I.K., B.W. and K.B. managed the mouse colony. D.J.C., T.S., D.A., J.S. and A.J.S. wrote the paper. All authors discussed and commented on the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to A.J.S. (silvaa@mednet.ucla.edu) or P.G. (pgolshani@mednet.ucla.edu).

METHODS

Subjects. All experimental protocols were approved by the Chancellor's Animal Research Committee of the University of California, Los Angeles, in accordance with NIH guidelines. Adult C57Bl/6NTac, C57Bl/6NTac \times 129S6/SvEvTac and C57Bl/6NIA male mice were singly housed on a 12 h light/dark cycle. Young adult mice were 3–6 months old, and aged adult mice were 14–18 months old. TetTag mice were generated by crossing transgenic mice that express a histone 2B–GFP fusion protein controlled by the tetO promoter (strain Tg(tetO-HIST1H2BJ/GFP) 47Efu/J; stock number 005104; Jackson Laboratory) with mice that express tetracycline-transactivator (tTA) protein under control of the *c-fos* (also known as *Fos*) promoter. TetTag mice were maintained in a C57Bl/6N background. Mice were born and raised on doxycycline (dox) chow (40 mg kg⁻¹) to prevent GFP expression before experimental manipulations. To open the window for activity-dependent labelling, dox chow was replaced with regular chow for 3 days before the start of an experiment. Expression of new GFP was shut off by administration of high dox chow (1 g kg⁻¹). Memory linking (transfer of fear and enhancement) experiments were conducted with both C57Bl/6NTac \times 129S6/SvEvTac and C57Bl/6NIA mice.

Viral construct. AAV1.Syn.GCaMP6f.WPRE.SV40 virus (titre: 4.65×10^{13} GC per ml) was purchased from Penn Vector Core. The hM3Dq vector was derived from the CaMK2a.hM4Di.T2A.EGFP/CREB plasmid³⁰. The hM4Di.T2A.EGFP/CREB in that plasmid was replaced by hM3Dq.T2A.EGFP/dTomato. The HA-tagged hM3Dq and dTomato-tagged EGFP are expressed under the CaMK2a promoter and cloned on either side of a T2A self-processing viral peptide. Vesicular stomatitis-virus-G-protein-pseudotyped lentiviral vectors were produced by calcium-phosphate-mediated transient transfection of human embryonic kidney 293 T (HEK293T) cells, as previously described³⁰. Lentivirus vectors were titred on HEK293T cells based on EGFP expression (titre: 6×10^5 cells per ml).

Surgery. Mice were anaesthetized with 1.5 to 2.0% isoflurane for surgical procedures and placed into a stereotaxic frame (David Kopf Instruments, Tujunga, CA). Lidocaine (2%; Akorn, Lake Forest, Illinois) was applied to the sterilized incision site as an analgesic, while subcutaneous saline injections were administered throughout each surgical procedure to prevent dehydration. In addition, carprofen (5 mg kg⁻¹) and dexamethasone (0.2 mg kg⁻¹) were administered both during surgery and for 7 days post-surgery with amoxicillin.

For calcium imaging experiments, mice underwent two separate surgical procedures. First, mice were unilaterally microinjected with 500 nl of AAV1.Syn.GCaMP6f.WPRE.SV40 virus at 50 nl min⁻¹ into the dorsal CA1 using the stereotaxic coordinates: -2.1 mm posterior to bregma, 2.0 mm lateral to midline and -1.65 mm ventral to skull surface. Two weeks later, the microendoscope (a gradient refractive index lens) was implanted above the previous injection site. For the procedure, a 2.0 mm diameter circular craniotomy was centred 0.5 mm medial to the virus injection site. Artificial cerebrospinal fluid (ACSF) was repeatedly applied to the exposed tissue to prevent drying. The cortex directly below the craniotomy was aspirated with a 27-gauge blunt syringe needle attached to a vacuum pump. The microendoscope (0.25 pitch, 0.50 NA, 2.0 mm in diameter and 4.79 mm in length, Grintech GmbH) was slowly lowered with a stereotaxic arm above CA1 to a depth of 1.35 mm ventral to the surface of the skull at the most posterior point of the craniotomy. Next, a skull screw was used to anchor the microendoscope to the skull. Both the microendoscope and skull screw were fixed with cyanoacrylate and dental cement. Kwik-Sil (World Precision Instruments) covered the microendoscope. Two weeks later, a small plastic baseplate was cemented onto the animal's head atop the previously formed dental cement. Debris was removed from the exposed lens with double-distilled H₂O, lens paper and forceps. The microscope was placed on top of the baseplate and locked in a position in which the field of focus was in view, so that cells and visible landmarks, such as blood vessels, appeared sharp and in focus. Finally, a plastic cover was fit into the baseplate and secured by magnets.

For aged DREADD experiments, mice were bilaterally microinjected with 700 nl of lentivirus CaMK2.hM3Dq.T2A.EGFP/dTomato virus at 100 nl min⁻¹ into the dorsal CA1 using the stereotaxic coordinates: -1.80 mm posterior to bregma, ± 1.50 mm lateral to midline, -1.60 mm ventral to skull surface; -2.50 mm posterior to bregma, ± 2.00 mm lateral to midline, -1.70 mm ventral to skull surface.

Drug injections. Clozapine-*N*-oxide (CNO; Enzo Life Sciences) was made in a stock solution of 0.5 mg ml⁻¹ in DMSO and then diluted in saline to desired concentration. CNO was injected (i.p.) at a dose of 0.5 mg kg⁻¹ 45 min before behavioural manipulation. MK-801 (Sigma-Aldrich) was diluted in saline and injected (i.p.) at a dose of 0.1 mg ml⁻¹ 30 min before behavioural manipulation. Saline was used as the vehicle.

Behavioural procedures. Prior to all experiments, mice were handled for one minute in the vivarium each day for three days. Then, mice were habituated to transportation and external environmental cues by being carted out of the vivarium into the experimental rooms and handled for one minute in the experimental room each day for five days before the experiment. For within-subject experiments,

mice explored three different contexts, separated by 7 days or 5 hours. Exploration duration of each context was ten minutes (C57Bl/6NTac and C57Bl/6NIA strains) or five minutes (C57Bl/6NTac \times 129S6/SvEvTac strain). Contexts were counter-balanced. For between-subject experiments, mice explored two contexts separated by either 7 days or 5 h. The area of each context was approximately 800 cm². The shape (circular, triangular, square), scent (simple green, omega, alcohol), visual cues (white plastic walls/opaque textured flooring, black acrylic walls/white acrylic flooring, metal walls/metal grid flooring) were different for each context. For immediate shock³¹ (imm shock), mice were placed in the chamber with a baseline of 10 s (0.7 mA) (C57Bl/6NTac and C57Bl/6NIA strains) or 6 s (C57Bl/6NTac \times 129S6/SvEvTac strain) followed by a 2 s shock (0.7 mA, C57Bl/6NTac and C57Bl/6NIA strains; 0.5 mA, C57Bl/6NTac \times 129S6/SvEvTac strain). Thirty seconds after the shock, mice were placed back in their home cage. For context tests (cxt test), mice were returned to the designated context. For extinction (extinct) trials, mice were placed in a context for five minutes without shock. Freezing (the cessation of all movement except for respiration), was assessed via an automated scoring system (Med Associates) with 30 frames per second sampling; the mice needed to freeze continuously for at least one second before freezing could be counted^{32,33}. Experimental groups and contexts were counter-balanced across the within-subjects design. For between-subjects design, animals were randomly assigned to groups.

Integrated miniature microscope data acquisition and analyses. Digital imaging data was sent from the CMOS imaging sensor (Aptina, MT9V032) to custom data acquisition (DAQ) electronics and USB host controller (Cypress, CYUSB3013) over a lightweight, highly flexible cable. The electronics packaged the data to comply with the USB video class (UVC) protocol and then transmit the data over Super Speed USB to a PC running custom DAQ software. The DAQ software was written in C++ and uses Open Computer Vision (OpenCV) libraries for image acquisition. Images are acquired at 30 frames per second and recorded to uncompressed .avi files. The DAQ software simultaneously records animal behaviour, time stamping both video streams for offline alignment.

Our analysis suite, written in MATLAB, processes the raw videos and extracts relevant experimental information. Initial processing of calcium imaging data corrected column-wise ADC variation, removed small movement artefacts using an amplitude-based image registration algorithm, and calculated the mean fluorescence per pixel for conversion to $\Delta F/F$. A fully automated segmentation algorithm identified and segmented pixels of active cells. The algorithm steps through the recorded calcium imaging video detecting pixel locations of local maxima of fluorescence which met a minimum $\Delta F/F$ criteria. For each of these pixel locations, an iterative process was used to group together neighbouring pixels based on that pixel's fluorescence time trace (± 5 s window around local maxima of fluorescence event) correlation with the mean time trace of the pixels group in the previous iterative step. Pixels with high correlation (0.95) were added to the group and the process was repeated until the total number of pixels in the group no longer changed. Cells whose centres were within $7 \mu\text{m}$ of each other or whose pixels overlapped by at least 80% were merged together. Once cells were segmented, we extracted $\Delta F/F$ traces and removed crosstalk between neighbouring cells. Crosstalk was removed by first detecting calcium transients across all cells and then keeping only the largest event within a $30 \mu\text{m}$ radius of the cell they were associated with⁷. Calcium events were calculated by first filtering the $\Delta F/F$ (2-pole Butterworth low-pass filter: 0.3 Hz) to remove noise. Peaks in the filtered $\Delta F/F$ trace above $0.05 \Delta F/F$ were detected and a window was calculated from the onset of the peak to the return back to baseline. If this window was greater than one second, it was counted as an event. Recordings from multiple sessions of the same animal were aligned using the same amplitude-based registration algorithm used for within-session registration, except the algorithm was only applied to the mean frame from each session. Once two sessions were registered, cells across two sessions were matched to each other using a distance measure (centres within $5 \mu\text{m}$ of each other).

Code availability. The MATLAB analysis suite, as described above, is available for download at <http://www.miniscope.org>. This Wiki site is our open-source platform for sharing access to all of our associated software and hardware files for implementing our miniature microscope.

Confocal imaging and histological analysis. Forty-five minutes after exploration of a context, mice were transcardially perfused with 4% PFA, followed by 24 h post-fixation in the same solution. Free-floating $50\text{-}\mu\text{m}$ coronal sections were prepared using a vibratome. Sections were incubated in blocking solution containing 0.2% Triton X, 10% normal goat serum in 0.1 M phosphate buffer for at least 1 h at room temperature. Then the sections were incubated in the blocking solution with anti-EGR-1 rabbit primary antibody (Cell Signaling; 1:750 dilution for 24 h at 4°C). After a series of 0.1 M phosphate buffer washes, sections were stained using the same blocking solution as above and Alexa Fluor 568 goat anti-rabbit secondary antibody (Jackson Immuno Research; 1:500 dilution for 2 h at room temperature).

Finally, sections were stained with DAPI (Invitrogen; 1:1,000 dilution for 15 min) and mounted on slides.

Sections from -1.8 mm to -2.2 mm posterior to bregma were imaged at $20\times$ magnification using a Nikon C2 or A1 confocal microscope. All imaging was done using standardized laser settings, held constant for samples from the same experimental data set. Cells were manually counted by a blinded rater. Images were quantified from 1–4 sections per animal. The percentage of DAPI-labelled cells containing GFP, ZIF, or both was calculated for each image and then averaged to produce a single measurement for each animal. To normalize for chance, we subtracted chance $(\text{GFP}/\text{DAPI}) \times (\text{ZIF}/\text{DAPI}) \times 100$ from the observed overlap $(\text{GFP and ZIF})/\text{DAPI} \times 100$ and then divided by chance.

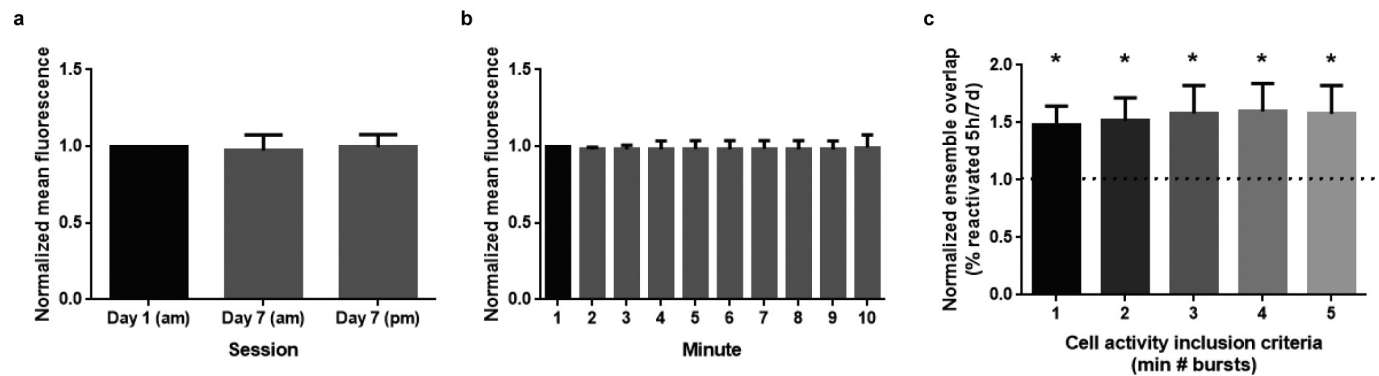
Electrophysiology. Mice were anaesthetized with a cocktail (3 ml kg^{-1}) containing ketamine (25 mg ml^{-1}), xylazine (1.3 mg ml^{-1}), and acepromazine (0.25 mg ml^{-1}) and perfused for 3 min with ice-cold, oxygenated, sucrose ACSF containing (in mM) 83 NaCl, 2.5 KCl, 3.3 MgSO_4 , 0.5 CaCl_2 , 1 NaH_2PO_4 , 26.2 NaHCO_3 , 22 glucose, and 72 sucrose ($\sim 315 \text{ mOsm}$, pH 7.4). The brain was rapidly dissected and $300\text{-}\mu\text{m}$ -thick coronal slices were collected and transferred to an interface chamber containing the same modified sucrose ACSF solution and incubated at 34°C for 30 min. Slices were then held at room temperature (23°C) in the interface chamber for at least 45 min before initiating recordings. Recordings were made in a submersion-type recording chamber and perfused with oxygenated ACSF containing (in mM) 119 NaCl, 2.5 KCl, 1.3 MgCl_2 , 2.5 CaCl_2 , 1.3 NaH_2PO_4 , 26.0 NaHCO_3 , 20 glucose ($\sim 295 \text{ mOsm}$) at 23°C at a rate of 1–2 ml per minute.

All recordings were performed within the CA1 region of the hippocampus. Neurons were selected based on emission spectra (GFP^+ or GFP^-), and were then visualized under infrared differential interference contrast video microscopy (Olympus BX-51 scope and Rolera XR digital camera). Whole-cell recordings were made at room temperature using pulled patch pipettes ($5\text{--}6 \text{ M}\Omega$) filled with

internal solution containing (in mM) 150 K-Gluconate, 1.5 MgCl_2 , 5.0 HEPES, 1 EGTA, 10 phosphocreatine, 2.0 ATP, and 0.3 GTP. Recordings were obtained using Multiclamp 700B patch amplifiers (Molecular Devices) and data analysed using pClamp 10 software (Molecular Devices). Data were acquired from cells requiring less than -100 pA to hold at a membrane potential of -70 mV . Current–spike relationship was determined with a series of depolarizing current steps applied for 500 ms in 10 pA increments at 5 s intervals.

Statistical analysis. GraphPad Prism version 6.00 (GraphPad Software, La Jolla, California, USA) was used for statistical analyses. Statistical significance was assessed by two-tailed paired Student's *t*-tests, two-tailed unpaired Student's *t*-tests, one-way ANOVA, or two-way ANOVA where appropriate. Significant effects or interactions were followed up with post hoc testing with the use of Fisher's least significant difference (LSD) where specified in the figure legends. Significance levels were set to $P=0.05$. Significance for comparisons: $*P<0.05$; $**P<0.01$; $***P<0.001$. Sample sizes were chosen on the basis of previous studies. No statistical methods were used to predetermine sample size. Data met assumptions of statistical tests, and variance was similar between groups for all metrics measured. The investigators were blinded to conditions and drugs during experiments and outcome assessment.

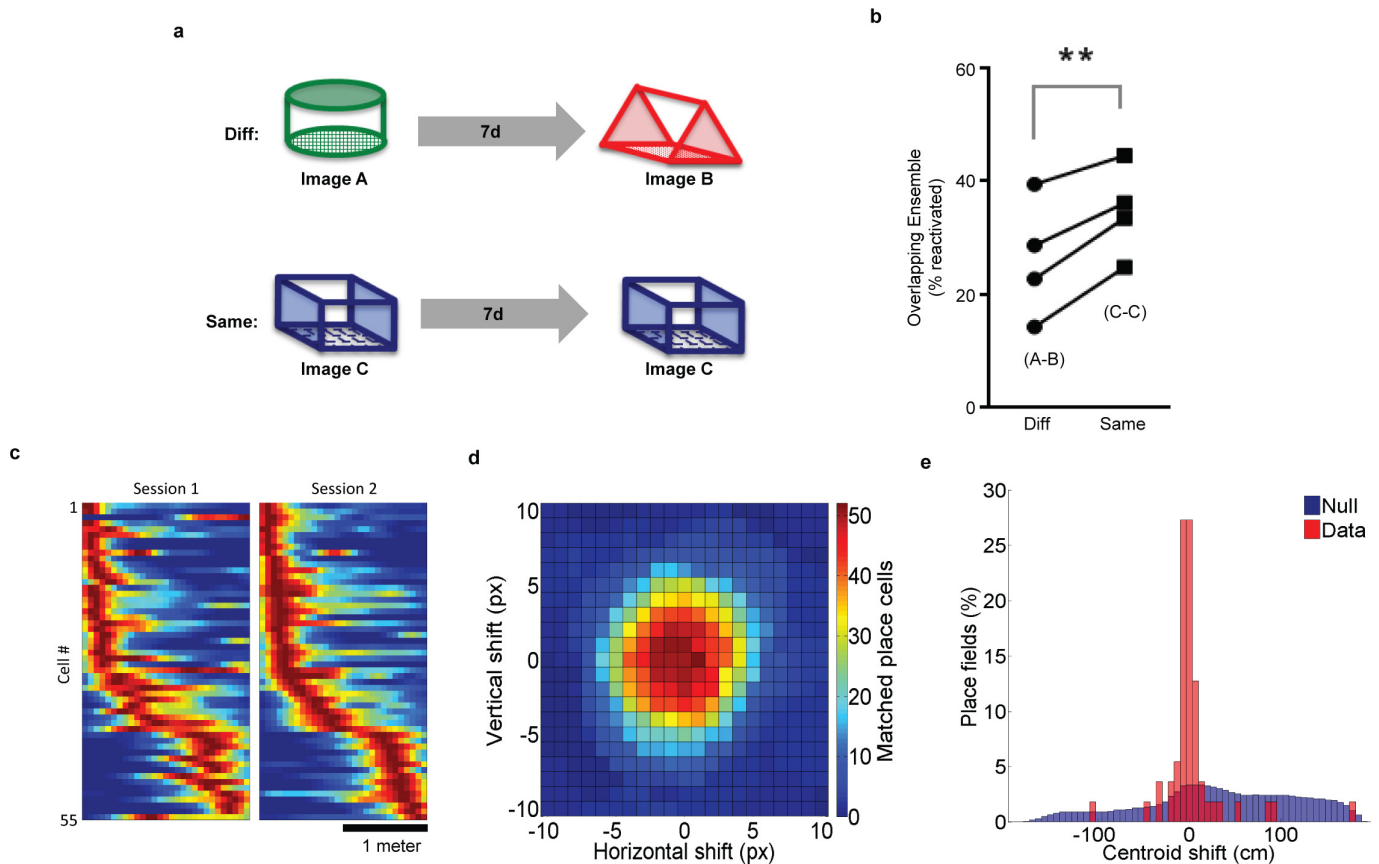
30. Sano, Y. *et al.* CREB regulates memory allocation in the insular cortex. *Curr. Biol.* **24**, 2833–2837 (2014).
31. Frankland, P. W. *et al.* Consolidation of CS and US representations in associative fear conditioning. *Hippocampus* **14**, 557–569 (2004).
32. Anagnostaras, S. G. *et al.* Automated assessment of pavlovian conditioned freezing and shock reactivity in mice using the video freeze system. *Front. Behav. Neurosci.* **4**, 158 (2010).
33. Cai, D. J., Shuman, T., Gorman, M. R., Sage, J. R. & Anagnostaras, S. G. Sleep selectively enhances hippocampus-dependent memory in mice. *Behav. Neurosci.* **123**, 713–719 (2009).



Extended Data Figure 1 | Stability of fluorescence and overlap.

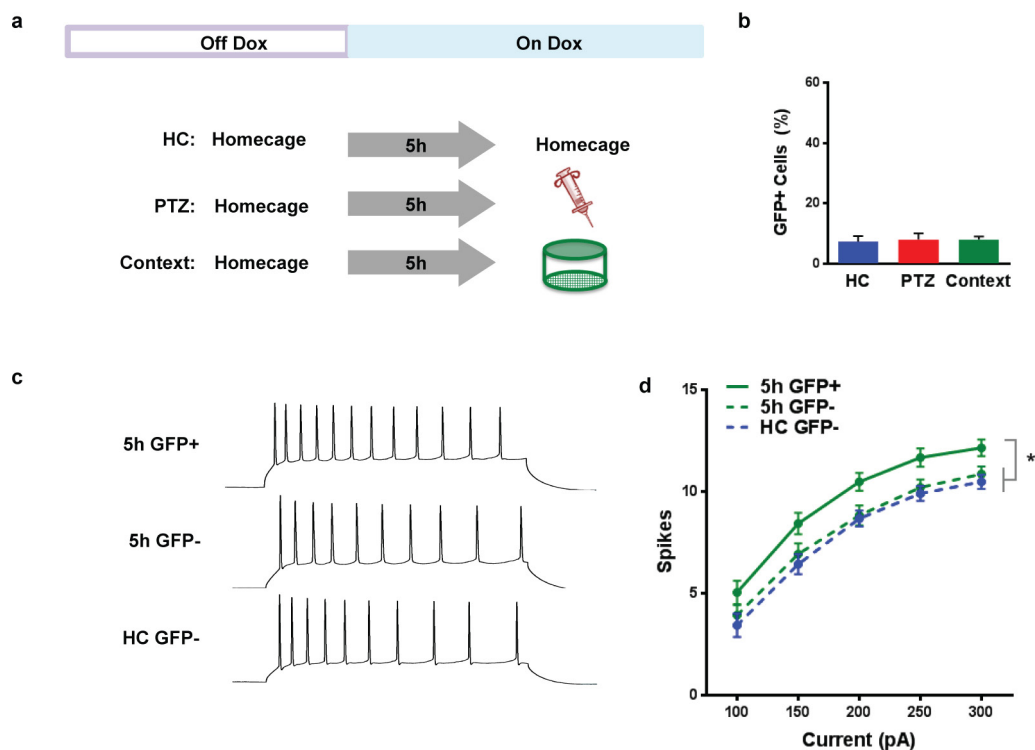
a, Average normalized mean fluorescence within session. There was no difference between the mean fluorescence across the 3 sessions (one-way repeated measures ANOVA, $F_{2,7} = 0.423$, not significant). **b**, Average normalized mean fluorescence within session. There was no difference between the mean fluorescence across a 10-min session (one-way repeated measures ANOVA, $F_{9,22} = 1.108$, not significant). Results show mean \pm s.d. **c**, Higher ensemble overlap with 5 h interval than 7 days. Normalized ensemble overlap is calculated as the ensemble overlap between contexts separated by 5 h divided by the ensemble overlap between contexts

separated by 7 days. A normalized overlap value of 1 signifies that there is no difference between the overlap at 5 h and 7 days. The minimum number of calcium events required from each cell for the cell to be considered 'active' (inclusion criteria) was systematically increased and the ratio of the ensemble overlap for the different context was calculated. For all inclusion criteria, there is higher ensemble overlap with a 5 h, rather than 7 day, interval (one-sample t -test against 1, (1) $t_7 = 3.00$, $P = 0.02$, (2) $t_7 = 2.57$, $P = 0.04$, (3) $t_7 = 2.42$, $P = 0.04$, (4) $t_7 = 2.50$, $P = 0.04$, (5) $t_7 = 2.32$, $P = 0.05$). Results show mean \pm s.e.m.



Extended Data Figure 2 | Neural ensembles of environments are reliably reactivated at recall of an open field and linear track. **a**, Experimental design. Mice were imaged while exploring contexts A and B separated by 7 days and imaged while exploring contexts C and C separated by 7 days. **b**, There was a higher percentage of cells reactivated when animals explored the same context (C–C) than when animals explored different contexts (A–B) (paired t -test, $t_3 = 6.305$, $P = 0.0081$, $n = 4$ mice). **c**, Mice were trained to run on a 2-m linear track with the miniature microscope for water rewards. Mice were trained 3 days a week for 3 weeks with a delay interval of 2–3 days between each session. Place fields were calculated by deconvolving calcium $\Delta F/F$ traces with an exponential to extract approximate spike times. Spikes that remained after crosstalk removal were included for analysis. Animal position was extracted using an automated LED tracking algorithm. A speed threshold (3 cm s^{-1}) was applied to both the animal position and extracted spike timing and the resulting data was spatially binned (6.5-cm bins). Spatial firing rates were calculated by dividing the binned spike counts by the binned occupancy

and smoothing with a Gaussian filter ($\sigma = 6.5 \text{ cm}$). Cells which showed consistent spatial firing modulation on at least three trials, with all other trials showing no bursting activity, were considered as place cells. Normalized spatial firing rates of all matched cells independently meeting the place cell criteria for both days. The data are pooled across 3 mice and include both motion directions. Place fields are ordered by centroid location on session 2. **d**, A shift of the image registration between sessions results in a decrease in matched place cells. A translational shift both horizontally and vertically was applied to the image registration transformation used in A. Cells were then matched across days and those which met our place cell criteria were kept. The heat map shows the count of matched place cells with a centroid shift of the place field that is less than 33 cm. Optimum matching of cells occurred within a 1-pixel translation of the calculated alignment transformation. **e**, Distribution of centroid shifts of place fields shown in A compared to the null hypothesis that the cell matching between sessions matches random cells.

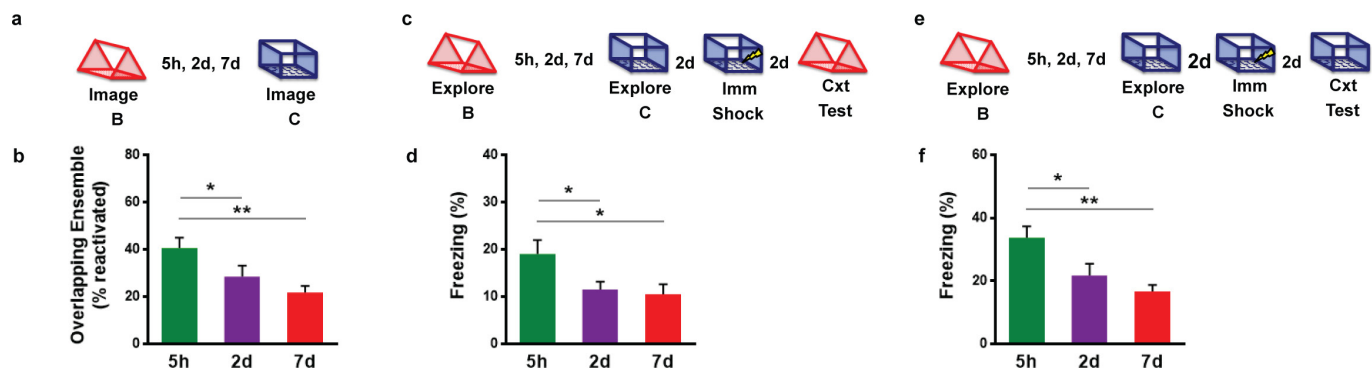


Extended Data Figure 3 | Five hours after exploration of a context, GFP expression is shut off by doxycycline and excitability is increased.

a, Experimental design. Mice were removed from low levels of dox (40 mg kg^{-1}) and given regular chow for 3 days to open up the GFP tagging window. After receiving administration of high dox (1 g kg^{-1}) for 5 h, mice were injected with 30 mg kg^{-1} of pentylenetetrazole (PTZ), exposed to a novel context or left in their home cage (HC). An hour later, mice were transcardially perfused and processed for GFP expression.

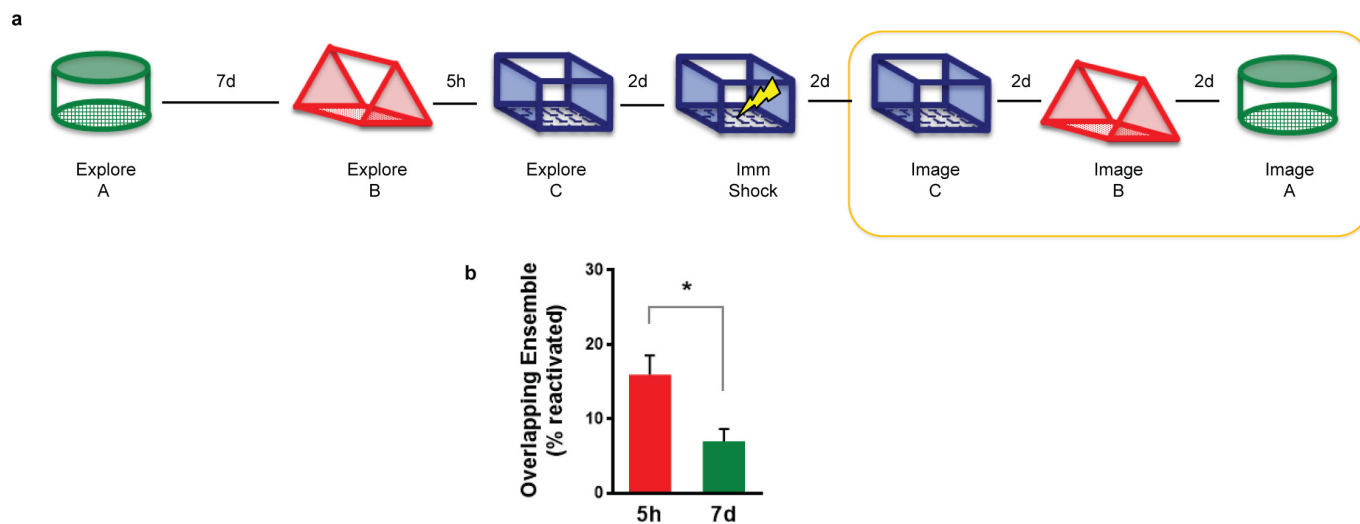
b, There was no difference in GFP expression between the three groups (one-way ANOVA, $F_{2,5} = 0.04$, not significant, $n = 3, 3, 2$ mice),

demonstrating that 5 h was enough time for dox (1 g kg^{-1}) to suppress expression of new GFP. **c**, To test excitability learning-related excitability changes, mice explored a novel context and then were administered high dox to shut off new GFP. Five hours later, mice were euthanized for *in vitro* slice physiology. **d**, A two-way repeated measures ANOVA (group \times current step) had a significant main effect of group ($F_{2,68} = 4.20$, $P < 0.05$, $n = 21, 29, 21$ cells). The 5 h GFP⁺ group had more spikes than the 5 h GFP⁻ group ($t_{68} = 2.31$, $P < 0.05$) and home cage GFP⁻ ($t_{68} = 2.72$, $P < 0.05$). There was no difference between the 5 h GFP⁻ and home cage GFP⁻ groups ($t_{68} = 0.61$, not significant). Results show mean \pm s.e.m.

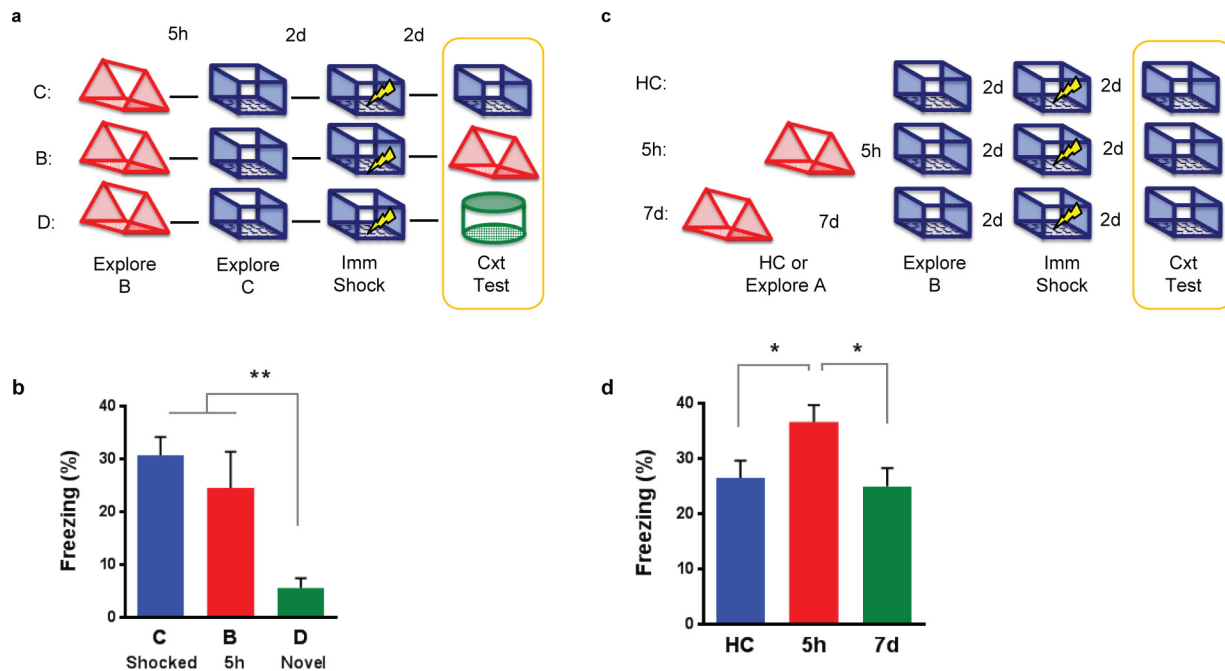


Extended Data Figure 4 | Time course for neuronal overlap and behavioural linking. **a**, Design for Ca^{2+} imaging of neuronal overlap experiment. **b**, There was a significant difference in overlap across groups (one-way repeated measures ANOVA, $F_{2,12} = 12.43$, $P = 0.002$, $n = 7$ mice). There was more overlap at 5 h than 2 days ($t_{12} = 3.03$, $P = 0.01$) and 7 days ($t_{12} = 4.72$, $P = 0.0005$). **c**, Design for transfer of fear experiment. **d**, There was a significant difference in freezing across groups (one-way ANOVA,

$F_{2,43} = 3.55$, $P = 0.04$, $n = 20, 14, 12$ mice). There was more freezing at 5 h than 2 days ($t_{43} = 2.13$, $P = 0.04$) and 7 days ($t_{43} = 2.31$, $P = 0.03$). **e**, Design for enhancement experiment. **f**, There was a significant difference in freezing across groups (one-way ANOVA, $F_{2,45} = 6.38$, $P = 0.004$, $n = 22, 14, 12$ mice). There was more freezing at 5 h than 2 days ($t_{45} = 2.45$, $P = 0.02$) and 7 days ($t_{45} = 3.32$, $P = 0.002$). Results show mean \pm s.e.m.

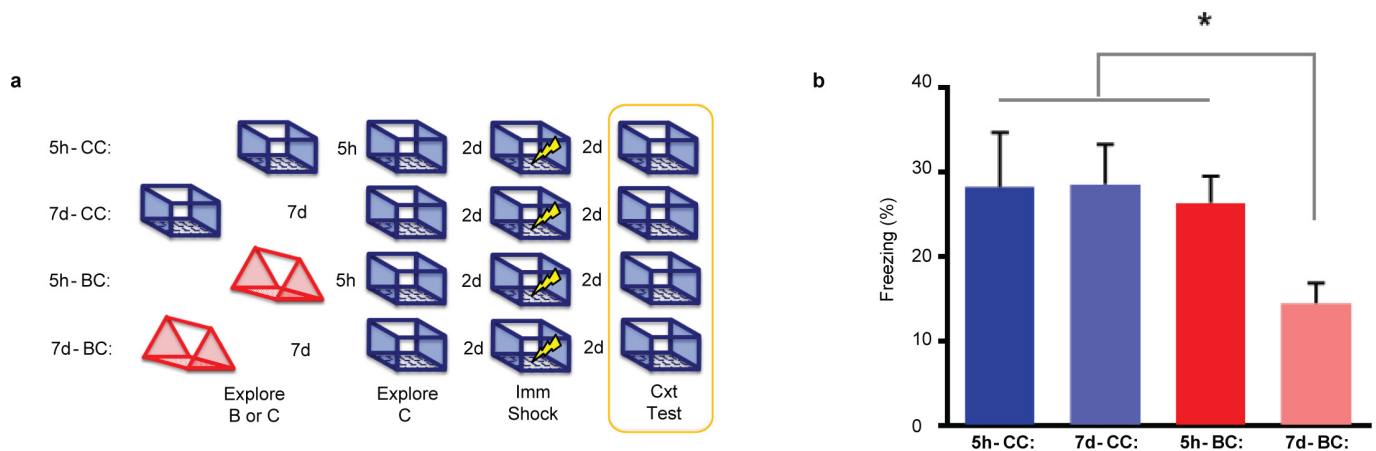


Extended Data Figure 5 | Calcium imaging during retrieval. **a**, Design for Ca^{2+} imaging of neuronal overlap at retrieval. Order of contexts during retrieval was counterbalanced. **b**, There was higher overlap of the neuronal ensemble at 5 h than 7 days (paired t -test, $t_7 = 2.55$, $P = 0.04$, $n = 8$ mice). Results show mean \pm s.e.m.



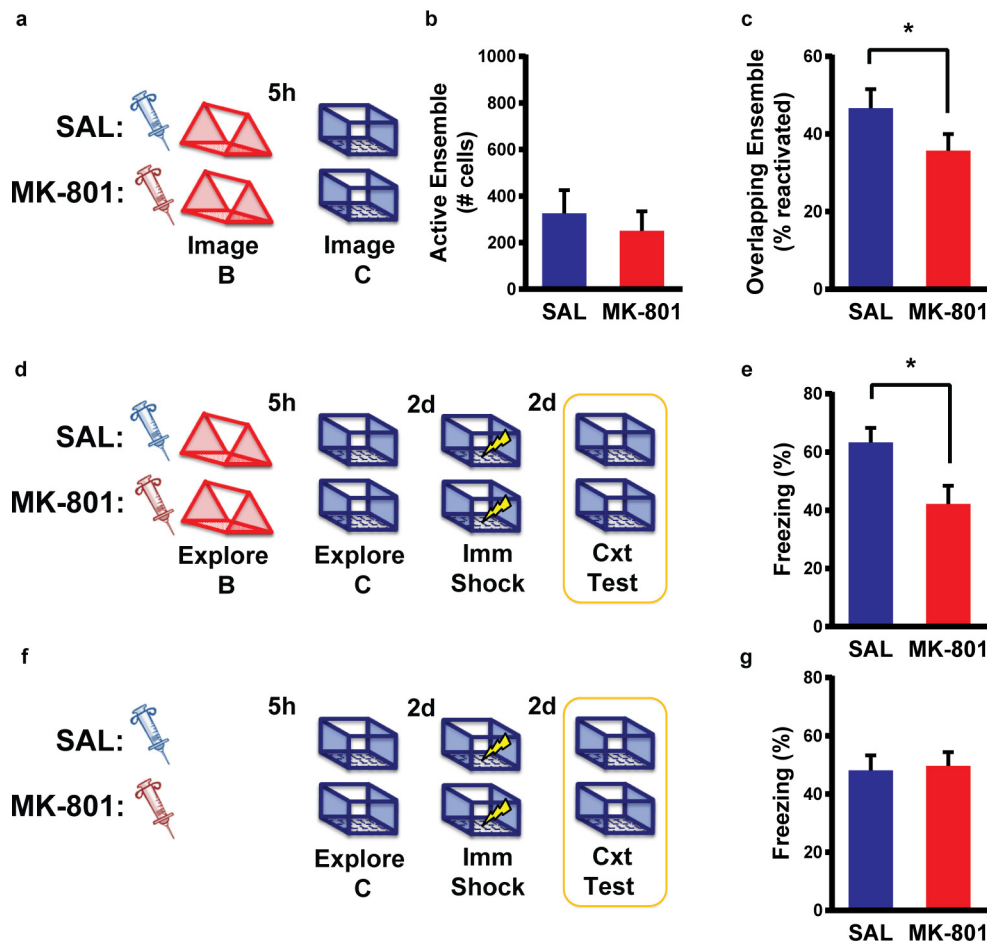
Extended Data Figure 6 | Replication of memory linking experiments in young (3–6 months old) C57Bl/6NIA mice. **a**, Design for transfer of fear experiment. **b**, There was a significant difference in freezing across the groups (one-way ANOVA, $F_{2,20} = 9.49$, $P = 0.001$, $n = 8, 7, 8$ mice). There was no difference between freezing levels in context C or B ($t_{20} = 0.99$, not significant). Animals had less freezing in context D than C ($t_{20} = 4.19$, $P = 0.0004$) and B ($t_{20} = 3.06$, $P = 0.006$). **c**, Design for enhancement

experiment. **d**, There was a significant difference in freezing (one-way ANOVA, $F_{2,46} = 4.071$, $P = 0.023$, $n = 16, 17, 16$ mice). The 5 h group had more freezing than the home cage (HC) group ($t_{46} = 2.72$, $P = 0.0278$) and 7 day group ($t_{46} = 2.612$, $P = 0.012$). There was no difference between home cage or 7 day groups ($t_{46} = 0.335$, not significant). Results show mean \pm s.e.m.



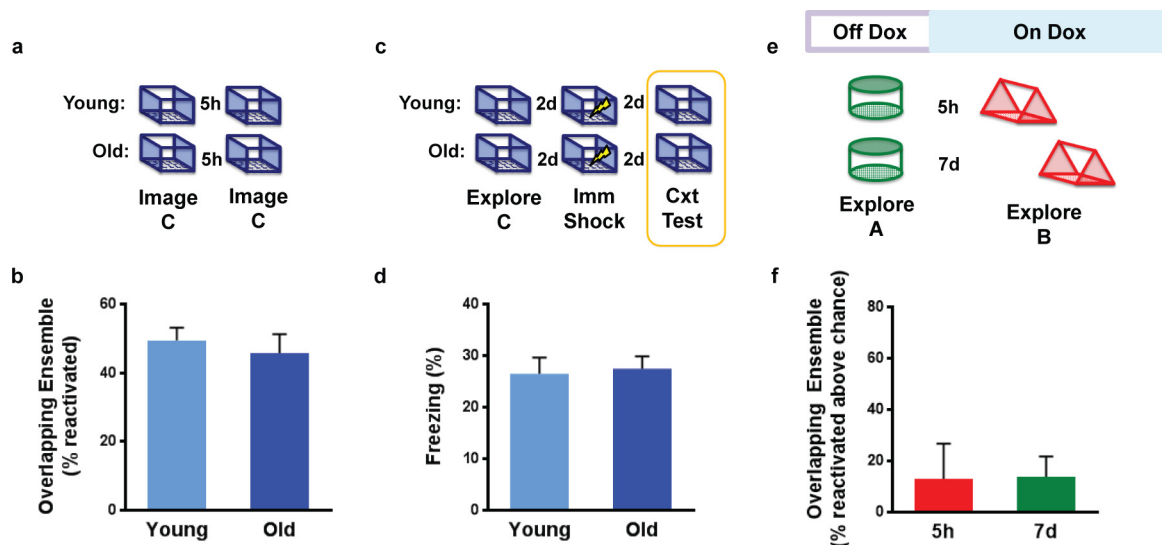
Extended Data Figure 7 | Exploring the same context twice enhances memory regardless of time. **a**, Experimental design. **b**, There was a significant difference in freezing (one-way ANOVA, $F_{3,44} = 2.92$, $P = 0.04$, $n = 10, 11, 13, 14$ mice). Consistent with the prior experiment, there was more freezing in the 5 h BC than the 7 day BC group ($t_{44} = 2.19$, $P < 0.05$).

The 7 day BC group also had more freezing than the 5 h CC ($t_{44} = 2.35$, $P < 0.05$) and 7 day CC ($t_{44} = 2.48$, $P < 0.05$) groups, however there were no difference between the 5 h CC and 7 day CC ($t_{44} = 0.06$, not significant) and 5 h CC and 5 h BC ($t_{44} = 0.31$, not significant) groups. Results show mean \pm s.e.m.



Extended Data Figure 8 | NMDA receptor activity is required for overlap of neural ensembles and behavioural enhancement. **a**, Design for Ca^{2+} imaging of neuronal overlap with MK-801 or saline. **b**, There was no difference in the number of cells active during exploration of the first context between saline-injected (SAL) and MK-801 groups (unpaired t -test, $t_6 = 0.58$, not significant, $n = 4, 4$). **c**, There was lower overlap of the neuronal ensemble in the MK-801 group than in the SAL group

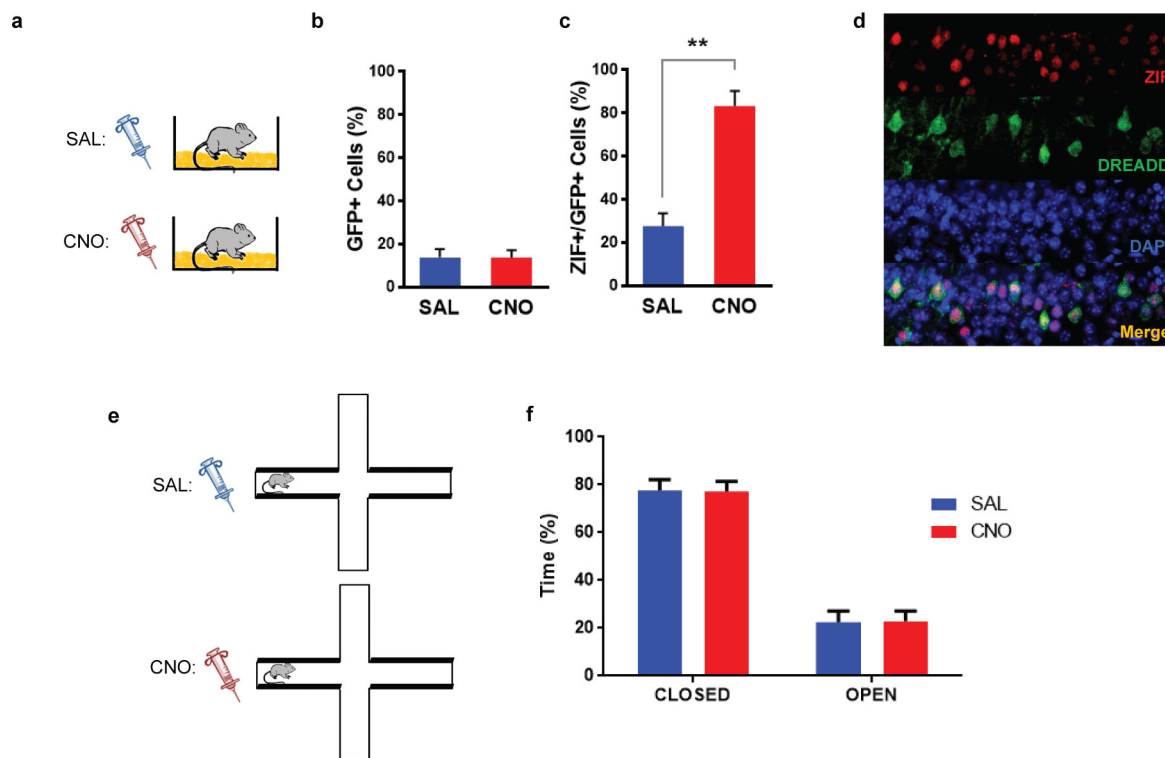
(paired t -test, $t_3 = 3.45$, $P = 0.04$, $n = 4$ mice). **d**, Design for behavioural enhancement experiment. **e**, There was lower freezing in the MK-801 than in the SAL group (unpaired t -test, $t_{22} = 2.65$, $P = 0.015$, $n = 12, 12$ mice). **f**, Design for behavioural control experiment. **g**, There was no difference in freezing between SAL and MK-801 groups (unpaired t -test, $t_{22} = 0.22$, not significant, $n = 12, 12$ mice). Results show mean \pm s.e.m.



Extended Data Figure 9 | Control experiments for aged mice. **a**, Design for experiment of recall for single contextual experience. **b**, There was no difference in reactivation of cells between young and old mice during recall (unpaired t -test, $t_6 = 0.59$, not significant, $n = 4$, 4 mice). **c**, Design for experiment with single context pre-exposure in young and old mice. **d**, There was no difference in freezing behaviour to exposures of a single

context (unpaired t -test, $t_{29} = 0.24$, not significant, $n = 16$, 15 mice).

e, Design for replication of TetTag experiment in old mice. **f**, There was no difference in the levels of overlapping ensembles between the 5 h and 7 day groups (unpaired t -test, $t_6 = 0.06$, not significant, $n = 3$, 5 mice). Results show mean \pm s.e.m.



Extended Data Figure 10 | CNO activates cells with DREADD

receptors and does not increase anxiety in aged mice. **a**, Mice infected with DREADD virus in CA1 were injected with saline (SAL) or clozapine-*N*-oxide (CNO) and then euthanized 90 min post-injection for immunofluorescence staining. **b**, There was no difference in the percentage of DREADD-positive cells (labelled with GFP) between SAL and CNO groups (unpaired *t*-test, $t_7 = 0.01$, not significant, $n = 3, 6$ mice). **c**, DREADD-positive cells (labelled with GFP) had more ZIF when

injected with CNO than SAL (unpaired *t*-test, $t_7 = 5.08$, $P = 0.02$).

d, Representative examples of ZIF, DREADD, DAPI as well as merged images of CA1. **e**, Design for elevated plus maze experiment in aged mice with DREADD virus. **f**, A two-way ANOVA showed no main effect of injection ($F_{1,9} = 0.75$, not significant, $n = 6, 5$ mice) and a significant main effect of arms ($F_{1,9} = 71.03$, $P < 0.0001$). There was no significant interaction between injection and arms ($F_{1,9} = 0.003$, not significant). Results show mean \pm s.e.m.

Pitx2 promotes heart repair by activating the antioxidant response after cardiac injury

Ge Tao¹, Peter C. Kahr¹, Yuka Morikawa², Min Zhang¹, Mahdis Rahmani², Todd R. Heallen², Lele Li¹, Zhao Sun³, Eric N. Olson⁴, Brad A. Amendt³ & James F. Martin^{1,2,5,6}

Myocardial infarction results in compromised myocardial function and heart failure owing to insufficient cardiomyocyte self-renewal¹. Unlike many vertebrates, mammalian hearts have only a transient neonatal renewal capacity². Reactivating primitive reparative ability in the mature mammalian heart requires knowledge of the mechanisms that promote early heart repair. By testing an established Hippo-deficient heart regeneration mouse model for factors that promote renewal, here we show that the expression of *Pitx2* is induced in injured, Hippo-deficient ventricles. *Pitx2*-deficient neonatal mouse hearts failed to repair after apex resection, whereas adult mouse cardiomyocytes with *Pitx2* gain-of-function efficiently regenerated after myocardial infarction. Genomic analyses indicated that *Pitx2* activated genes encoding electron transport chain components and reactive oxygen species scavengers. A subset of *Pitx2* target genes was cooperatively regulated with the Hippo pathway effector Yap. Furthermore, Nrf2, a regulator of the antioxidant response³, directly regulated the expression and subcellular localization of *Pitx2*. *Pitx2* mutant myocardium had increased levels of reactive oxygen species, while antioxidant supplementation suppressed the *Pitx2* loss-of-function phenotype. These findings reveal a genetic pathway activated by tissue damage that is essential for cardiac repair.

We used immunofluorescence staining to look for developmental transcription factors that were upregulated in regenerating Hippo-deficient hearts⁴. Paired-like homeodomain transcription factor 2 (*Pitx2*) was enriched in border zone ventricular cardiomyocyte nuclei of adult Hippo-deficient mouse hearts after myocardial infarction (Fig. 1a–c). *Pitx2* encodes three isoforms (*Pitx2a*, *Pitx2b* and *Pitx2c*). It functions in left–right asymmetric organ development⁵ and is mutated in Rieger syndrome, which is characterized by craniofacial, umbilical and cardiac abnormalities⁶. Notably, *Pitx2* deficiency results in

predisposition to the common human arrhythmia atrial fibrillation^{7,8}. *Pitx2c* is the major isoform expressed in heart.

Available RNA-sequencing (RNA-seq) data indicated that *Pitx2* transcripts in cardiomyocytes dropped postnatally⁹ (Fig. 1d), and western blot analysis revealed *Pitx2* protein induction after injury during regenerative stages (Fig. 1e). Consistent with reduced *Pitx2* expression in adult hearts, active histone marks at the *Pitx2* locus were reduced in

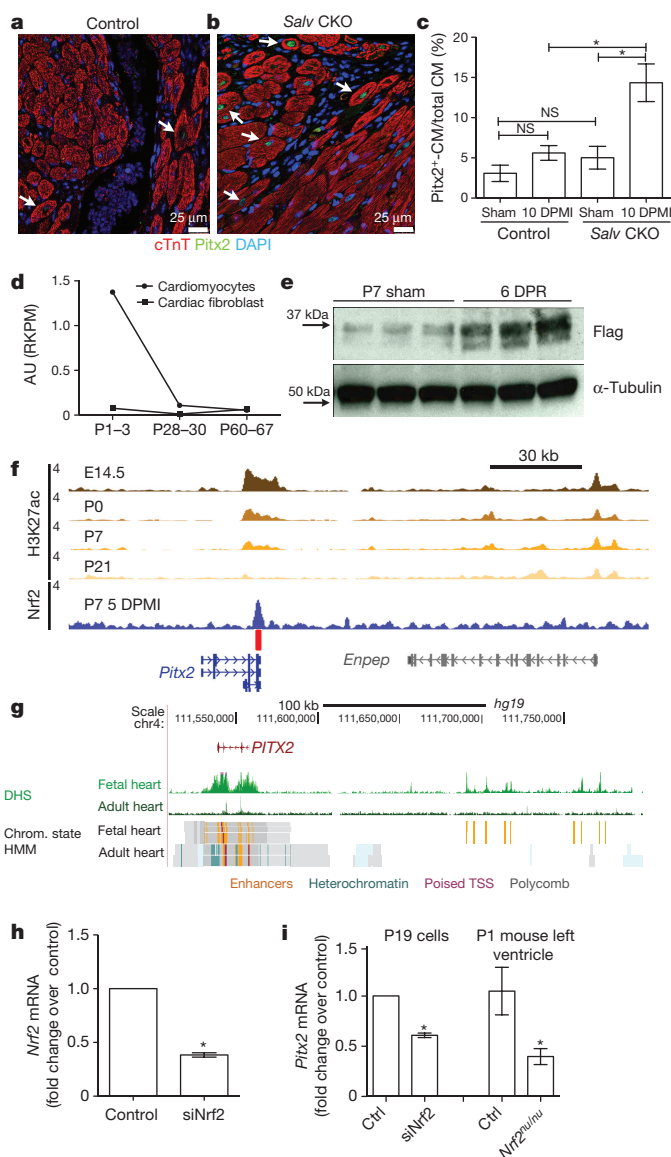


Figure 1 | *Pitx2* is induced in injured myocardium. **a–c**, Border zone of *Salv* KO (**b**) and control (**a**) hearts stained for *Pitx2* (green), cTnT (red), and DAPI (blue), at 10 days after myocardial infarction (DPPI, days post-myocardial infarction), with the *Pitx2*⁺ cardiomyocyte (CM) ratio quantified in **c**; *n* = 5 mice per group. **d**, *Pitx2* expression shown by RNA-seq. AU, arbitrary units; RPKM, reads per kilobase of transcript per million mapped reads. **e**, Western blot of Flag and α-tubulin in 6 DPPI *Pitx2*^{flag} ventricles, resected at P1, compared to sham; *n* = 3 mice per group. **f**, Nrf2 protein directly binds to the *Pitx2* enhancer after LAD-O. The heart-specific enhancers are marked by H3K27ac ChIP-seq. Red bar denotes the Nrf2-binding element. **g**, DHS-seq and chromatin state hidden Markov model (Chrom. state HMM) tracks of fetal and adult human heart tissue. Orange indicates active enhancer regions. TSS, transcription start sites. **h**, qPCR shows short interfering RNA (siRNA) knockdown of *Nrf2* (siNrf2) in P19 cells; *n* = 4 biological replicates. **i**, qPCR of *Pitx2* in P19 cells with siRNA targeting *Nrf2*, and in *Nrf2*^{mut/mut} heart, compared to controls; *n* = 4 biological replicates. Data are mean ± s.e.m. **P* < 0.05, one-way analysis of variance (ANOVA) plus Bonferroni post-test (**c**), and Mann–Whitney test (**h**, **i**) (see Methods). NS, not significant.

¹Department of Molecular Physiology and Biophysics, Baylor College of Medicine, One Baylor Plaza, Houston, Texas 77030, USA. ²Texas Heart Institute, Houston, Texas 77030, USA. ³Department of Anatomy and Cell Biology and the Craniofacial Anomalies Research Center, The University of Iowa, Iowa City, Iowa 52242, USA. ⁴Department of Molecular Biology and Hamon Center for Regenerative Science and Medicine, University of Texas Southwestern Medical Center, Dallas, Texas 75390-9148, USA. ⁵Program in Developmental Biology, Baylor College of Medicine, Houston, Texas 77030, USA. ⁶Cardiovascular Research Institute, Baylor College of Medicine, Houston, Texas 77030, USA.

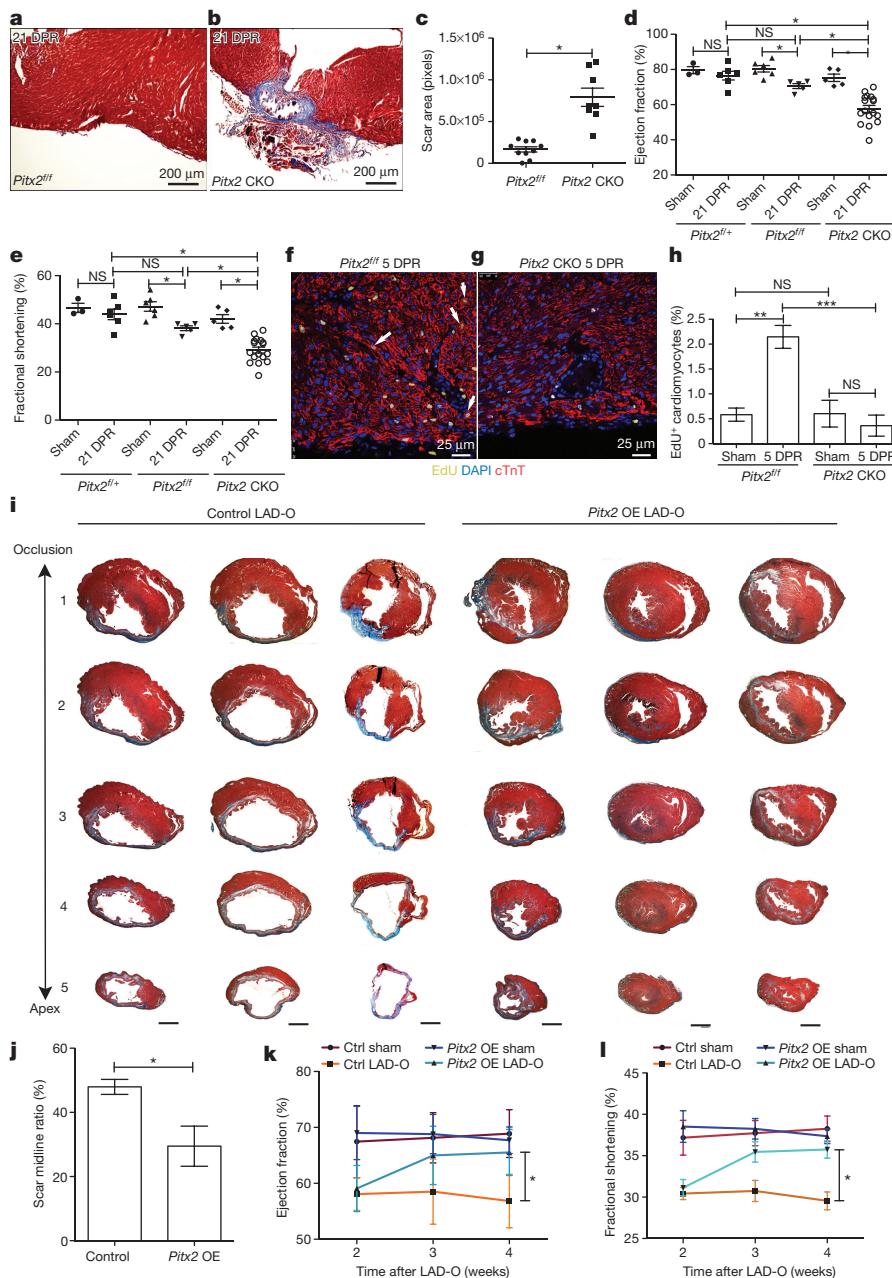


Figure 2 | *Pitx2* is required and sufficient to promote myocardial regeneration.

a–c, Trichrome-stained *Pitx2*^{fl/fl} (a) and *Pitx2* CKO (b) apex at 21 DPR, with scar size quantified in c (see Methods for *n*). **d, e**, Echocardiography shows the ejection fraction (d) and fractional shortening (e) at 21 DPR (see Methods for *n*). **f–h**, 5 DPR *Pitx2*^{fl/fl} (f) and *Pitx2* CKO (g) apical sections stained for EdU (yellow), cTnT (red), and DAPI (blue). Arrows, EdU⁺ cardiomyocytes. The cardiomyocyte proliferative ratio was quantified in h; *n* = 4 mice per group. **i**, Serial transverse heart sections at 5 weeks after LAD-O, performed in 8-week-old mice. OE, overexpressing. **j**, Percentage of fibrotic left ventricular myocardium quantified at 5 weeks after LAD-O (see Methods for *n*). Scale bar, 1 mm. **k, l**, Ejection fraction (k) and fractional shortening (l) of LAD-O and sham hearts (see Methods for *n*). Data are mean \pm s.e.m. **P* < 0.05, one-way ANOVA plus Bonferroni post-test (d, e), and Mann-Whitney test (c, h, j–l).

adult mouse hearts¹⁰ (Fig. 1f, g). Available DNase I hypersensitive sites (DHSs) coupled with high-throughput DNA sequencing (DHS-seq) data revealed that Nrf2 (also known as Nfe2l2) binding elements were enriched at the *Pitx2* locus (data not shown). To evaluate whether Nrf2 activated *Pitx2* after injury, we performed an Nrf2 chromatin immunoprecipitation followed by sequencing (ChIP-seq) experiment on hearts 4 days after left anterior descending artery occlusion (LAD-O) in postnatal day 2 (P2) mice, and discovered Nrf2 binding at the *Pitx2* locus (Fig. 1f). Nrf2 knockdown in P19 cells and Nrf2 loss-of-function in mice resulted in decreased *Pitx2* mRNA expression, and supports the conclusion that Nrf2 directly regulates *Pitx2* after tissue injury (Fig. 1h, i).

We determined whether *Pitx2*, similarly to *Yap1*, was required for neonatal heart regeneration². Using Cre recombinase driven by the muscle creatine kinase (*MCK*, also known as *Ckm*) gene (*MCK*^{cre})¹¹, we inactivated *Pitx2* in cardiomyocytes and performed P1 apex resection. While control hearts regenerated as expected, *MCK*^{cre};*Pitx2*^{fl/fl} (*Pitx2* conditional knockout (CKO)) hearts had increased scarring and reduced function (Fig. 2a–e). We injured *Pitx2* mutant hearts by LAD-O at P1, and used both *MCK*^{cre} and *Mhc*^{cre-Ert} to inactivate *Pitx2* in myocardium. *Pitx2* mutants failed to repair after LAD-O (Extended Data Fig. 1).

We examined cardiomyocyte proliferation in the P1 apex resection mouse model at 5 days post-resection (DPR) by pulse-labelling and immunofluorescence of 5-ethynyl-2'-deoxyuridine (EdU). In *Pitx2*^{fl/fl} controls, injury induced a threefold increase in EdU-positive cardiomyocytes compared to sham that was absent in *Pitx2* CKO mice after injury, supporting the hypothesis that *Pitx2*, like *Yap1*, is essential for neonatal heart regeneration by promoting proliferation and injury resistance (Fig. 2f–h).

To investigate whether *Pitx2* is sufficient for adult cardiomyocyte repair, we generated *Pitx2*^{Gof}, a Cre-activated *Pitx2c* gain-of-function transgenic line (Extended Data Fig. 2a). Immunoblotting and quantitative PCR (qPCR) showed increased *Pitx2c* levels in *Mhc6*^{cre-Ert};*Pitx2*^{Gof} (*Pitx2*-overexpressing) ventricles (Extended Data Fig. 2b–d). LAD-O performed in 8-week-old mice after tamoxifen administration revealed that *Pitx2*-overexpressing hearts had reduced scar size⁴ (Fig. 2i, j). Heart morphology was comparable between controls (*Myh6*^{cre-Ert/+}) and *Pitx2*-overexpressing mice after sham surgery (Extended Data Fig. 2e–g). Two weeks after LAD-O, both *Pitx2*-overexpressing and controls showed decreased ejection fraction and fractional shortening, however, *Pitx2*-overexpressing mice had functional recovery at

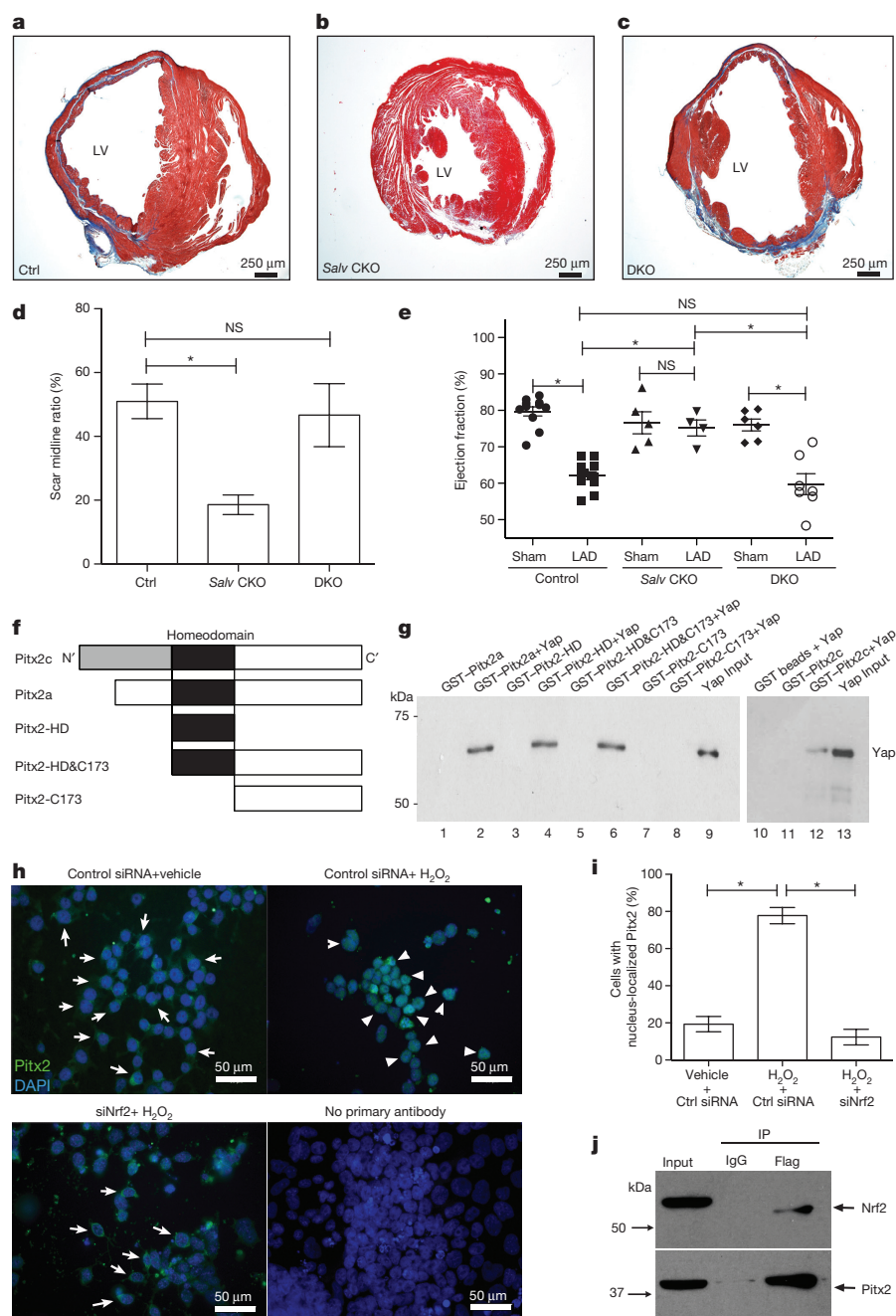


Figure 3 | Pitx2 interacts with Yap in regenerating hearts, and its nuclear shuttling requires Nrf2. **a–d**, Trichrome-stained control (*Salv*^{f/f};*Pitx2*^{f/f}, *n* = 5) (**a**), *Salv* CKO (*n* = 5) (**b**) and double knockout (DKO, *n* = 4) (**c**) sections at 28 days after LAD-O in P8 mice, with scar size quantification (**d**). **e**, Echocardiography shows the ejection fraction (see Methods for *n*). **f**, Diagram of GST-Pitx2 constructs. **g**, GST-Pitx2 pull-down assay. Yap was detected by western blotting. **h**, **i**, Immunofluorescent staining of Pitx2 (green) and DAPI (blue) in P19 cells after vehicle or 300 μ M H₂O₂ treatment, with control siRNA or siRNA targeting *Nrf2* (siNrf2). Arrows, cytoplasmic staining; arrowheads, nuclear staining. The ratio of cells with nucleus-localized Pitx2 compared to total cell number is quantified in **i**; 3 technical replicates per experiment, repeated 3 times. **j**, Co-immunoprecipitation of Flag in 5 DPR *Pitx2*^{flag} ventricles, resected at P1, with blotting of Nrf2 and Pitx2. Mean \pm s.e.m. **P* < 0.05, one-way ANOVA plus Bonferroni post-test (**e**) and Mann–Whitney test (**d**, **i**).

3 and 4 weeks after LAD-O (Fig. 2k, l). Non-regenerative stage P8 apex resections revealed that hearts from *Pitx2*-overexpressing mice had reduced scarring (Extended Data Fig. 2h–j) and improved function at 28 DPR compared to controls (Extended Data Fig. 2k, l). EdU incorporation at 8 DPR showed increased cardiomyocyte S-phase entry in *Pitx2*-overexpressing mice hearts (Extended Data Fig. 2m–o).

Because *Pitx2* was upregulated in Hippo-deficient hearts, we tested whether *Pitx2* was required for Hippo-deficient cardiomyocyte renewal. *Salv* (also known as *Sav1*) CKO hearts regenerate efficiently after myocardial infarction⁴. However, *Salv* CKO hearts that were also *Pitx2* mutant (double knockout) failed to regenerate (Fig. 3a–c). Twenty-eight days after P8 LAD-O, double knockout hearts had a larger scar and compromised ejection fraction⁴ (Fig. 3d, e). Apex resection in non-regenerative P8 hearts also revealed the requirement for *Pitx2* function in *Salv* CKO cardiomyocyte renewal (Extended Data Fig. 3).

Available genomic footprinting data from cardiac DHS data sets can uncover sequence-specific transcription factor–DNA interactions in an unbiased fashion. Motifs for Pitx2 and Tead, the Yap DNA-binding

partner, were highly enriched in fetal heart footprints and often found in close proximity (Extended Data Fig. 4a, b). Genomic regions containing *Pitx2* or *Tead* motifs were enriched for histone 3 Lys4 methylation (H3K4me1) chromatin marks, indicating that Pitx2- or Tead-binding regions were transcriptionally active. Regions containing both *Pitx2* and *Tead* motifs showed globally increased transcriptional activity compared to regions containing only *Pitx2* motifs (Extended Data Fig. 4c, d).

The co-occurrence of transcription factor binding motifs often indicates transcription factor interactions. We tested whether Pitx2 was a Yap binding partner using purified glutathione S-transferase (GST) fusion proteins. *In vitro* binding assays with Pitx2 fusion peptides and full-length Yap revealed that Yap bound the Pitx2 homeodomain (Fig. 3f, g, Extended Data Fig. 5a, b). We uncovered an *in vivo* interaction between endogenous Pitx2 and Yap using co-immunoprecipitation of endogenous cardiac proteins (Extended Data Fig. 5c). The *Pitx2*^{flag} allele, previously generated by gene targeting in mouse embryonic stem (ES) cells, expresses endogenous levels of Flag-epitope-tagged Pitx2 from the *Pitx2* locus⁷.

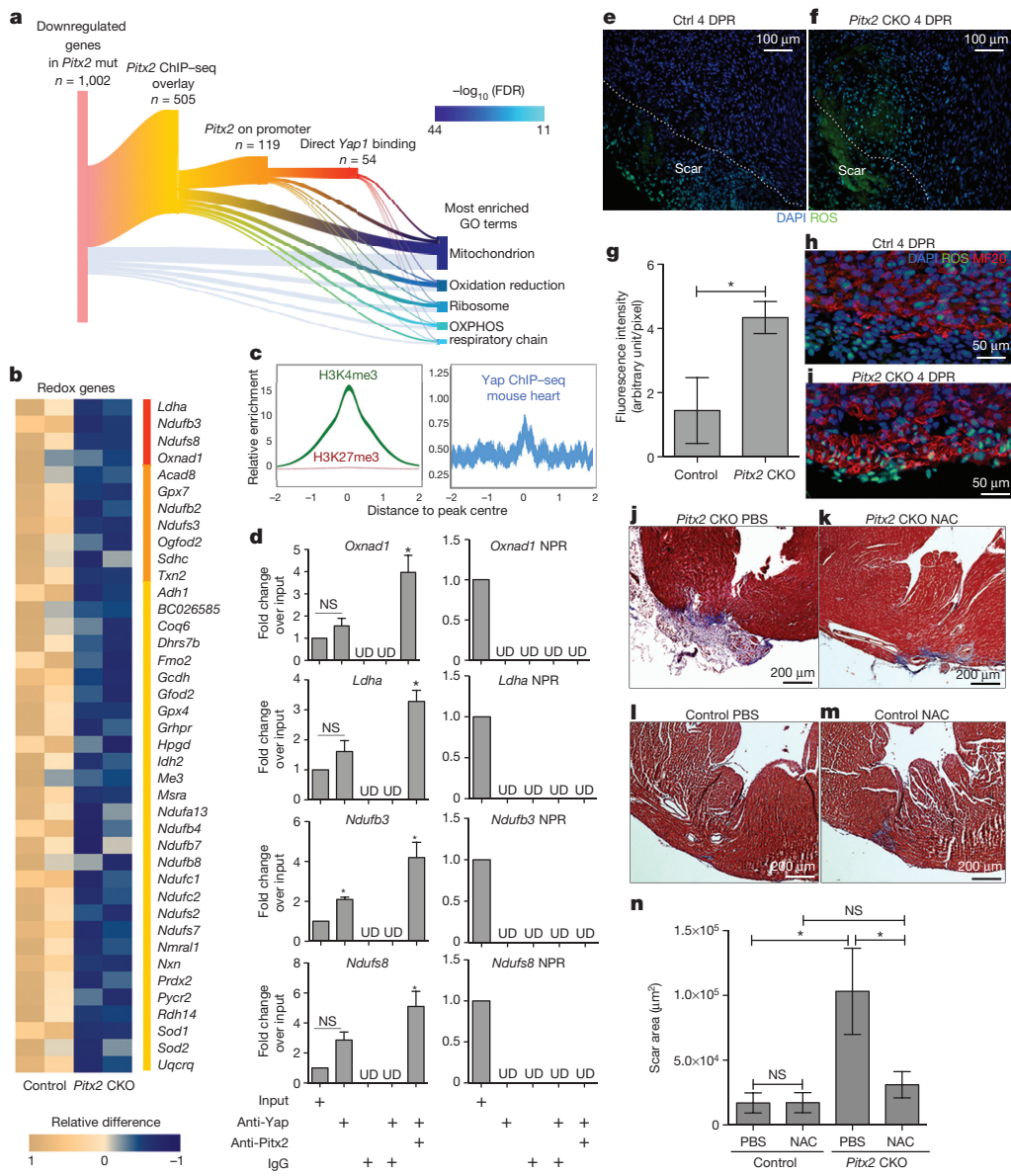


Figure 4 | *Pitx2* regulates redox balance in neonatal myocardium. **a**, Sankey diagram shows direct target genes of *Pitx2* from overlaying ChIP-seq and RNA-seq profiles.

A subset of genes is further branched according to the promoter-binding activity of *Pitx2* and *Yap1*. GO analysis is performed on 1,002 downregulated genes in *Pitx2* CKO ventricles at 5 days after P1 resection. The GO terms are listed by significance in descending order. The branches indicate the source of genes for each term. OXPHOS, oxidative phosphorylation. **b**, Heat map highlights mitochondrial genes directly targeted by *Pitx2*. Red bar, genes co-regulated by *Pitx2* and *Yap1*; orange, direct binding of *Pitx2* on promoters. **c**, Heart-specific H3K4me3 ChIP-seq and *Yap* ChIP-seq read distributions within a 2-kb promoter region of *Pitx2* direct target genes. The width of the curve indicates the 95% confidence interval. NPR, non-peak region; UD, undetermined. **d**, ChIP-re-ChIP showing co-occupancy of *Pitx2* and *Yap* at the regulatory regions of common target genes (in **b**, red bar); $n = 3$ biological replicates. **e–i**, ROS staining (green) of apical border zone in *Pitx2* CKO (**f**, **i**) and control (**e**, **h**) mice, with fluorescence intensity quantified in **g**. MF20 (myosin heavy chain antibody), red; DAPI, blue; $n = 4$ mice per group. **j–n**, Trichrome at 21 DPR showed apical scarring of *Pitx2* CKO (**j**, **k**) and control (**l**, **m**) hearts treated with PBS (**j**, **l**) or NAC (**k**, **m**), with scar area quantified in **n**; $n = 5$ mice per group. Mean \pm s.e.m. * $P < 0.05$, one-way ANOVA plus Bonferroni post-test (**n**), Mann–Whitney test (**g**) and Wilcoxon signed-rank test (**d**).

Immunofluorescence analysis showed widespread distribution of *Pitx2* in P19 cells, and a cytoplasm-to-nucleus translocation after hydrogen peroxide (H_2O_2) treatment (Fig. 3h, Extended Data Fig. 6a), similar to the Nrf2 response to oxidative stress³ (Extended Data Fig. 6b). *Pitx2* nuclear translocation after H_2O_2 treatment depended on Nrf2 activity (Fig. 3h, i). By contrast, Nrf2 nuclear translocation after H_2O_2 treatment was intact in *Pitx2*-null P19 cells indicating that *Pitx2* was dispensable for the Nrf2 response to reactive oxygen species (ROS) (Extended Data Fig. 6b, c). We found that *Pitx2* interacts with Nrf2 in heart extracts, expressing endogenous protein levels (Fig. 3j). Co-immunoprecipitation experiments using nuclear–cytoplasmic fractionation of P19 cells and analysing endogenous proteins indicated that *Pitx2* binding to Nrf2 in the nucleus was increased after H_2O_2 treatment (Extended Data Fig. 6d, e). We also found less nuclear *Pitx2* in *Nrf2*-mutant hearts after P1 apex resection (Extended Data Fig. 6f–h).

To solidify the connection between Nrf2, *Pitx2* and *Yap*, we tested whether Nrf2 was required for neonatal regeneration, as is the case for *Pitx2* and *Yap*. Myocardial infarction in P2 mice revealed that *Nrf2*-null hearts were unable to regenerate, indicating that induction of the antioxidant response is required for regeneration¹² (Extended Data Fig. 7). Notably, *Pitx2*-overexpressing mice that were heterozygous for the *Nrf2*-null allele (*Nrf2*^{nu/nu}) failed to regenerate, suggesting that *Pitx2* promotes the antioxidant response. It is also possible that Nrf2

is downstream of *Pitx2* in certain contexts. We also made *Pitx2*-overexpressing mice that were heterozygous for a floxed *Yap1* allele (*Yap1*^{f/+})¹³. Reducing *Yap1* dosage compromised *Pitx2*-overexpressing heart regeneration in a P8 mouse resection model (Extended Data Fig. 8a–e).

To investigate *Pitx2* target genes induced by injury, we collected P1 resected ventricles from *Pitx2*^{fl/fl} and *Pitx2* CKO hearts at 5 DPR and performed RNA-seq (Extended Data Fig. 9a–d). We identified 1,002 downregulated genes in *Pitx2* mutants (false discovery rate (FDR) ≤ 0.1 , fold change ≥ 1.5). There was extensive overlap between upregulated genes after apex resection in controls and downregulated genes in 5-DPR *Pitx2* CKO hearts, indicating that in the absence of *Pitx2*, a set of stress response genes, including oxidative stress response genes, fails to be activated (Extended Data Fig. 9a–d).

We examined the response of *Pitx2* and antioxidant scavenger genes to H_2O_2 in *Pitx2*-null (*Pitx2*^{nu/nu}) ES cells since *Pitx2* has been implicated in the ROS response in skeletal muscle^{6,14}. After H_2O_2 treatment, qPCR showed increased *Pitx2*, *Gpx1*, *Mt1* and *Mt2* expression in wild-type ES cells, but not in *Pitx2*^{nu/nu} ES cells (Extended Data Fig. 9f), supporting a crucial role for *Pitx2* in the response to ROS. While ES cells had low endogenous *Pitx2* levels, the mouse P19 embryonic carcinoma cell line expressed readily detectable *Pitx2*, primarily the *Pitx2c* isoform. H_2O_2 -treated P19 cells increased the

expression levels of *Pitx2* and its target gene in a dose-dependent manner (Extended Data Fig. 9g, h).

To identify *Pitx2* target genes, we performed P1 apex resection and ChIP-seq on 5-DPR *Pitx2*^{flg} ventricles (Extended Data Fig. 9e). Overlay of downregulated genes from *Pitx2* CKO RNA-seq with *Pitx2*-binding peaks from ChIP-seq revealed 505 direct *Pitx2* targets. Gene Ontology (GO) analysis revealed enrichment in mitochondria, oxidation-reduction, ribosome and respiratory chain genes (Fig. 4a, b).

Among *Pitx2* targets were genes that protect the cell from increased ROS, such as the superoxide dismutase genes *Sod1* and *Sod2*, which reduce superoxide to H₂O₂, and the glutathione peroxidase gene *Gpx4*, which removes H₂O₂, and *Prdx2* (ref. 15) (Fig. 4b). *Pitx2* regulates electron transport chain complex I components including *Ndufb3*, *Ndufb4* and *Ndufb7*, and complex IV component *Cox7c* (Fig. 4b). Defective complex I in human patients increases ROS sensitivity¹⁶. *Pitx2* regulated 21.5% of its direct target genes through promoter binding, as revealed by enrichment of H3K4me3 chromatin marks for active promoters in *Pitx2* peaks (Fig. 4c; 119 out of 505 direct targets). 8-week-old mouse heart H3K4me3 chromatin marks are enriched in the centre of *Pitx2*-binding sites, supporting the hypothesis that *Pitx2* promotes transcriptional activation¹⁷.

To determine whether *Pitx2* and Yap regulate common target genes, we performed Yap ChIP-seq on ventricles 5 days after LAD-O in P2 mice, and found Yap-binding sites enriched in nearly half of the *Pitx2*-targeted gene promoters (54 out of 119; Fig. 4a, c). Comparison of *Pitx2* ChIP-seq, our Yap ChIP-seq and available Yap ChIP-seq^{18–21} data revealed four redox genes bound by both *Pitx2* and Yap. ChIP-re-ChIP assay from heart extracts revealed *Pitx2* and Yap were concurrently resident on these genes, indicating that Yap and *Pitx2* cooperatively activate the transcriptional response to oxidative stress (Fig. 4d).

To investigate ROS activity in *Pitx2*^{flg} and *Pitx2* CKO apical border zones at 4 DPR, tissue sections were incubated with ROS-detecting reagent. *Pitx2* CKO hearts had increased ROS in both cardiomyocytes and non-myocytes (Fig. 4e–i). To determine whether increased ROS contributed to scarring in 21 DPR *Pitx2* CKO hearts, we administered N-acetyl-L-cysteine (NAC) in *Pitx2* CKO neonates after apex resection. Daily NAC injections until 10 DPR decreased scar size in *Pitx2* CKO hearts (Fig. 4j–n).

Increased ROS is a natural response to cardiac injury including ischaemic damage^{22,23} (Extended Data Fig. 10). During the postnatal transition from glycolytic to oxidative metabolism, ROS is increased in the heart and inhibits cardiomyocyte regeneration²². In regenerative-stage hearts, *Pitx2* promotes regeneration by inhibiting ROS. Injury induces *Pitx2* expression and activity through Nrf2-activated *Pitx2* transcription and nuclear shuttling. In turn, *Pitx2* activates ROS scavengers, protecting cells from oxidative damage, and electron transport chain components. It is also possible that Nrf2 also acts downstream of *Pitx2* in some contexts. Thus, *Pitx2* is essential for the cardiomyocyte response to injury and may prevent cell death.

We uncovered a *Pitx2*–Yap interaction important for Hippo-deficient cardiac regeneration. *Pitx2* binds Yap and cooperatively activates genes maintaining redox balance. *Pitx2* gain-of-function, sufficient to bestow reparative capacity in adult cardiomyocytes, is repressed by *Yap1* heterozygosity. This suggests that *Pitx2* recruits Yap to target genes even when Hippo is active and the pool of nuclear Yap is relatively low. This mechanism may work in parallel with other mechanisms by which Yap protects the cell from ROS²⁴.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 8 June 2015; accepted 29 March 2016.

Published online 25 May 2016.

- Xin, M., Olson, E. N. & Bassel-Duby, R. Mending broken hearts: cardiac development as a basis for adult heart regeneration and repair. *Nature Rev. Mol. Cell Biol.* **14**, 529–541 (2013).

- Porrello, E. R. *et al.* Transient regenerative potential of the neonatal mouse heart. *Science* **331**, 1078–1080 (2011).
- Itoh, K. *et al.* Keap1 represses nuclear activation of antioxidant responsive elements by Nrf2 through binding to the amino-terminal Neh2 domain. *Genes Dev.* **13**, 76–86 (1999).
- Heallen, T. *et al.* Hippo signaling impedes adult heart regeneration. *Development* **140**, 4683–4690 (2013).
- Lu, M. F., Pressman, C., Dyer, R., Johnson, R. L. & Martin, J. F. Function of Rieger syndrome gene in left-right asymmetry and craniofacial development. *Nature* **401**, 276–278 (1999).
- Semina, E. V. *et al.* Cloning and characterization of a novel *bicoid*-related homeobox transcription factor gene, *RIEG*, involved in Rieger syndrome. *Nature Genet.* **14**, 392–399 (1996).
- Wang, J. *et al.* *Pitx2* prevents susceptibility to atrial arrhythmias by inhibiting left-sided pacemaker specification. *Proc. Natl Acad. Sci. USA* **107**, 9753–9758 (2010).
- Kirchhof, P. *et al.* PITX2c is expressed in the adult left atrium, and reducing *Pitx2c* expression promotes atrial fibrillation inducibility and complex changes in gene expression. *Circ. Cardiovasc. Genet.* **4**, 123–133 (2011).
- Giudice, J. *et al.* Alternative splicing regulates vesicular trafficking genes in cardiomyocytes during postnatal heart development. *Nature Commun.* **5**, 3603 (2014).
- Nord, A. S. *et al.* Rapid and pervasive changes in genome-wide enhancer usage during mammalian development. *Cell* **155**, 1521–1531 (2013).
- Bruning, J. C. *et al.* A muscle-specific insulin receptor knockout exhibits features of the metabolic syndrome of NIDDM without altering glucose tolerance. *Mol. Cell* **2**, 559–569 (1998).
- Chan, K., Lu, R., Chang, J. C. & Kan, Y. W. Nrf2, a member of the NFE2 family of transcription factors, is not essential for murine erythropoiesis, growth, and development. *Proc. Natl Acad. Sci. USA* **93**, 13943–13948 (1996).
- Xin, M. *et al.* Hippo pathway effector Yap promotes cardiac regeneration. *Proc. Natl Acad. Sci. USA* **110**, 13839–13844 (2013).
- L'Honoré, A. *et al.* Redox regulation by *Pitx2* and *Pitx3* is critical for fetal myogenesis. *Dev. Cell* **29**, 392–405 (2014).
- Dhalla, N. S., Temsah, R. M. & Netticadan, T. Role of oxidative stress in cardiovascular diseases. *J. Hypertens.* **18**, 655–673 (2000).
- Larsson, N. G. & Clayton, D. A. Molecular genetic aspects of human mitochondrial disorders. *Annu. Rev. Genet.* **29**, 151–178 (1995).
- Yue, F. *et al.* A comparative encyclopedia of DNA elements in the mouse genome. *Nature* **515**, 355–364 (2014).
- Zanconato, F. *et al.* Genome-wide association between YAP/TAZ/TEAD and AP-1 at enhancers drives oncogenic growth. *Nature Cell Biol.* **17**, 1218–1227 (2015).
- Galli, G. G. *et al.* YAP drives growth by controlling transcriptional pause release from dynamic enhancers. *Mol. Cell* **60**, 328–337 (2015).
- Stein, C. *et al.* YAP1 exerts its transcriptional control via TEAD-mediated activation of enhancers. *PLoS Genet.* **11**, e1005465 (2015).
- Morikawa, Y. *et al.* Actin cytoskeletal remodeling with protrusion formation is essential for heart regeneration in Hippo-deficient mice. *Sci. Signal.* **8**, ra41 (2015).
- Puente, B. N. *et al.* The oxygen-rich postnatal environment induces cardiomyocyte cell-cycle arrest through DNA damage response. *Cell* **157**, 565–579 (2014).
- Chouchani, E. T. *et al.* Ischaemic accumulation of succinate controls reperfusion injury through mitochondrial ROS. *Nature* **515**, 431–435 (2014).
- Shao, D. *et al.* A functional interaction between Hippo-YAP signalling and FoxO1 mediates the oxidative stress response. *Nature Commun.* **5**, 3315 (2014).

Supplementary Information is available in the online version of the paper.

Acknowledgements The project was supported in part by IDRC grant number 1U54 HD083092 from the Eunice Kennedy Shriver National Institute of Child Health & Human Development. This project was supported by the Mouse Phenotyping Core at Baylor College of Medicine with funding from the National Institutes of Health (NIH) (U54 HG006348). The project was also supported by grants from the NIH (DE 023177 and HL 118761 to J.F.M.; DE 13941 to B.A.A.; HL-077439, HL-111665, HL-093039, DK-099653 and U01-HL-100401 to E.N.O.), and the Vivian L. Smith Foundation (J.F.M.). J.F.M. was supported by Transatlantic Network of Excellence Award LeDucq Foundation Transatlantic Networks of Excellence in Cardiovascular Research 14CVD01: “Defining the genomic topology of atrial fibrillation”. E.N.O. was supported by Fondation Leducq Networks of Excellence, Cancer Prevention and Research Institute of Texas and the Robert A. Welch Foundation (grant 1-0025). G.T. was supported by American Heart Association (AHA) (13POST17040027). P.C.K. was supported by German Research Foundation (DFG) (KA4018/1-1).

Author Contributions J.F.M. and G.T. conceived the project and designed the experiments. G.T., P.C.K., Y.M., M.R., T.R.H. and L.L. performed experiments and analysed data. Z.S. and B.A.A. provided reagents and performed experiments. E.N.O. provided the transgenic animal model. M.Z. and G.T. performed bioinformatics and statistical analyses. J.F.M. supervised the project and analysed data. G.T. and J.F.M. wrote the manuscript.

Author Information The sequencing data set has been deposited in the NCBI Gene Expression Omnibus (GEO) under accession number GSE70413. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to J.F.M. (jfmartin@bcm.edu).

METHODS

Mouse alleles and transgenic lines. All animal protocols and procedures were approved by the Institutional Animal Care and Use Committee (IACUC) of Baylor College of Medicine, Houston, Texas 77030, USA. All surgeries and echocardiographic studies were carried out blinded from genotype of the mice. Littermate controls were used whenever possible. Both male and female mice were used. The *MCK^{cre}*, *Myh6-cre/Esr1* (*Mhc^{cre-Ert}*) transgenic line, floxed alleles for *ww45/salvador* (*Salv^{fl}*) and *Pitx2* (*Pitx2^{fl}*), and Flag-tagged *Pitx2* allele (*Pitx2^{flag}*) have been described previously^{7,25}. The *Pitx2^{Gof}* construct for overexpressing *Pitx2* was generated by introducing a 0.9 kb *Pitx2c* cDNA coding sequence into a CMV-CAG-loxP-eGFP-Stop-loxP-IRES-βGal expression vector²⁶, linearized construct was subjected to pronuclear injection. *MCK^{cre};Pitx2^{fl}* (*Pitx2* CKO), *Mhc^{cre-Ert};Salv^{fl}* (*Salv* CKO), *Mhc^{cre-Ert};Salv^{fl};Pitx2^{fl}* (DKO) and *Mhc^{cre-Ert};Pitx2^{Gof}* (*Pitx2*-overexpressing) mice were generated by cross breeding. After Cre-mediated recombination, the *Pitx2* floxed allele removes all *Pitx2* isoform function. DNA was extracted from tail biopsies for genotyping. The primers for *Pitx2^{Gof}* are: forward, 5'-CACATGAAGCAGCAGACTT-3'; reverse, 5'-TGCTCAGGTAGTGGTTGTCG-3'. *Nrf2^{nu/nu}* is available from the Jackson Laboratory (strain name B6.129X1-Nfe2l2^{tm1Ywk}/J, stock number 017009)¹². *Yap^{fl}* strain has been described previously²⁷.

Cardiac apex resection. Surgical resection of the heart apex was performed on P1 and P8 mice as described previously^{2,28}. For P8 surgery, tamoxifen was administered daily from P7–P10, by subcutaneous injection at a dosage of 40 mg kg⁻¹ (ref. 28). Vicryl sutures (6-0 absorbable) were used to close the thoracic cavity, and the entire procedure required approximately 12 minutes from the onset of hypothermia to recovery. Sham procedures excluded apex amputation. Mice were subjected to echocardiography and then euthanized at 21 DPR for P1 resection, or 28 DPR for P8 resection. Dissected hearts were processed for histology and immunohistochemistry. Fibrotic scar size was measured using ImageJ 1.43u (National Institutes of Health) and the *n* number for each group is as follows: Fig. 2, 10 for *Pitx2^{fl}*; 8 for *Pitx2* CKO; Supplementary Fig. 2, 10 for *Mhc^{cre-Ert}*, and 7 for *Pitx2*-overexpressing; Supplementary Fig. 3, 5 for *Salv^{fl}*; *Pitx2^{fl}*; 7 for *Salv* CKO; 3 for DKO; Supplementary Fig. 4, 8 for each genotype.

LAD-O. LAD-O in P8 mice was performed according to previous descriptions; tamoxifen was administered daily from P7–P10, by subcutaneous injection at a dosage of 40 mg kg⁻¹ (ref. 28). Nylon sutures (8-0 non-absorbable) were used to occlude the LAD. Proper occlusion was noted by blanching of the myocardium and during dissection 3–4 weeks after occlusion via visual inspection. Vicryl sutures (6-0 absorbable) were used to close the thoracic cavity, and the entire procedure required approximately 12 minutes from the onset of hypothermia to recovery. Sham procedures excluded placement of a suture around the LAD. Mice were subjected to echocardiography, and then euthanized at 3–4 weeks after occlusion. Hearts were processed for histology and immunohistochemistry. Automated fibrotic scar size was measured using image segmentation MIQuant, open source code for Matlab²⁹. The *n* number for each group is as follows: Fig. 3, 5 for *Salv^{fl}*; *Pitx2^{fl}*, 5 for *Salv* CKO, 4 for DKO. Alternatively, LAD-O was performed at P2, with minor modification from P1 apex resection and P8 LAD-O procedures described above, tamoxifen was administered daily from P2–P3 when needed³⁰. The *n* number for each group is as follows: Supplementary Fig. 1, 8 for *Pitx2^{fl}*, 7 for *MCK^{cre};Pitx2^{fl}*, 4 for *Mhc^{cre-Ert};Pitx2^{fl}*.

Adult LAD-O was performed as described for P8 with minor modifications²⁸. For *Pitx2*-overexpressing and control (*Mhc^{cre-Ert}*) mice, surgery was performed in 8-week-old mice, and tamoxifen was administered by intraperitoneal injection at three time points: 7 and 6 days before LAD-O and within 2 h after LAD-O, at a dosage of 40 mg kg⁻¹. Echocardiography was performed at 2, 3 and 4 weeks after LAD-O. The mice were then euthanized and hearts were subjected to histology. Automated fibrotic scar size was measured as described for P8 LAD-O. The *n* number for each group is as follows: Fig. 1, 5 for control and 5 for *Salv* CKO, 5 sham controls for each group; Fig. 2, 5 for *Mhc^{cre-Ert}* LAD-O, 8 for *Pitx2*-overexpressing LAD-O. 5 sham controls were used for each genotype.

Echocardiography. Echocardiography was performed in the Baylor College of Medicine Mouse Phenotyping Core using a VisualSonics 2100 system. Evaluation of ejection fraction and fractional shortening of apex resection model was performed as previously described^{2,28}. The *n* number for each group is as follows: Fig. 2, *Pitx2^{fl}*, 3 for sham, 6 for resection; *Pitx2^{fl}*, 6 for sham, 5 for resection; *Pitx2* CKO, 5 for sham, 17 for resection; adult LAD-O, control (*Mhc^{cre-Ert}*), 4 for sham, 5 for LAD-O; *Pitx2*-overexpressing, 5 for sham, 8 for LAD-O. Fig. 3, control (*Salv^{fl};Pitx2^{fl}*), 11 for sham, 12 for LAD-O; *Salv* CKO, 5 for sham, 4 for LAD-O; DKO, 6 for sham, 7 for LAD-O. Supplementary Fig. 1, *Pitx2^{fl}*, 7 for sham, 8 for LAD-O; *MCK^{cre};Pitx2^{fl}*, 5 for sham, 7 for LAD-O; *Mhc^{cre-Ert};Pitx2^{fl}*, 4 for sham, 4 for LAD-O. Supplementary Fig. 2, control (*Mhc^{cre-Ert}*), 6 for sham, 25 for resection; *Pitx2*-overexpressing, 5 for sham, 16 for resection. Supplementary Fig. 7, control (C57BL6), 3 for sham, 5 for LAD-O; *Nrf2^{nu/nu}*, 3 for sham, 5 for LAD-O.

EdU incorporation. For P6 and P16 mice, 0.25 mg of EdU was injected subcutaneously 7 h before collecting the hearts. After dissection, hearts were fixed with 10% neutral buffered formalin and processed for paraffin embedding. Seven-micrometre-thick tissue sections were prepared. EdU incorporation was detected using the Click-it EdU imaging kit (Life Technologies). Tissue slides were imaged with a Leica TCS SP5 confocal microscopy, and images were processed by Leica LAS AF software (Leica Microsystems).

Cardiomyocyte proliferation studies. To assess cardiomyocyte proliferation rates, 5 DPR *Pitx2* CKO and control (*Pitx2^{fl}*) mice and 8 DPR *Pitx2*-overexpressing and control (*Mhc^{cre-Ert}*) mice (as described earlier) were used. EdU labelling and detection were performed as described above. Mouse monoclonal anti-cTnT (1:200) (Thermo Scientific) was used to label cardiomyocytes. Images were acquired as described earlier. The cardiomyocyte proliferation rate was calculated by dividing the number of EdU-positive cardiomyocytes by the total number of cardiomyocytes in the field. Three comparable sections (every third section) from each heart were used.

NAC administration. NAC (PharmaGrade, A5099 Sigma-Aldrich) was solved in sterile PBS at a concentration of 10 mg ml⁻¹. After P1 apex resection, mice were weighed daily, and NAC solution was injected subcutaneously from 1 to 10 DPR at a dosage of 75 mg kg⁻¹. Three comparable sections (every third section) from each heart were used, and five hearts were used in each group in Fig. 4j–n.

Cell culture. P19 cells (ATCC CRL-1825) were cultured in αMEM medium (Mediatech, Corning), supplemented with 10% FBS (Gibco, Life Technologies) and 1% penicillin/streptomycin (HyClone Laboratories, Thermo Scientific). 0.25% trypsin was used for dissociating and splitting cells. H₂O₂ (Sigma-Aldrich) and doxorubicin (D-4000, LC Laboratories) were diluted in αMEM with 1% FBS and 1% penicillin/streptomycin at final concentrations of 300 μM and 0.5 μM, respectively. After 8 h of treatment, cells were collected and subjected to mRNA or protein extraction. The ES cells used in this study have been described previously⁶. Mycoplasma detection kit (B39030, <http://www.bioutil.com>) and MycoAlert kit (LT07-318, Lonza) were used, and no contamination was observed.

Transfection of siRNA in P19 cells. Lipofectamine RNAiMAX transfection reagent (ThermoFisher Scientific) was used to deliver siRNA targeting *Nrf2* into P19 cells following the manufacturer's guideline. The siRNA oligonucleotides were pre-designed DsiRNA Duplex from Integrated DNA Technologies. Oligonucleotide sequences: antisense, rGrArUrGrArArUrCrArArUrCrCrArUrGrUrCrCrUrGrCrUrG; sense, rGrCrArGrGrArCrArUrGrGrArUrUrGrArUrGrArCrATC.

Generation of P19 *Pitx2* knockout cell line using CRISPR-Cas9 technique. Lentivirus expressing *Cas9* was used to transduce P19 cells for generating stable cell line expressing *Cas9* (J.F.M. *et al.*, unpublished data). Guides targeting exons 5 and 6 of the *Pitx2* locus were designed using Optimized CRISPR Design (<http://crispr.mit.edu>, F. Zhang laboratory, MIT 2015). Guides sequences used: upstream, 5'-CACCGAATGAGGATGTGGGCGCCG, 3'-AAACCGGCGCCACATCCTCATTC; downstream, 5'-CACCGTGTCCCTA TAAACGTACGG, 3'-AAACCCGTACGTTTATAGGGACAC. Guides were inserted into pSpCas9(BB)-2A-GFP (PX458) (F. Zhang, Addgene plasmid 48138)³¹. *Cas9*-expressing P19 cells were transfected with both guide plasmid simultaneously, using Lipofectamine 2000 Transfection Reagent (Thermo Scientific) according to the manufacturer's manual. GFP-positive cells were sorted using a BD FACSAria cell sorter. Single-cell clones were expanded and genotyped using the following primers: (1) 494 bp for wild type, undetectable for knockout: forward: 5'-GCACACACCCACACTTTCAC-3', reverse: 5'-CTTCCACCCACCACTCCTAC-3'; (2) 272 bp for wild type, undetectable for knockout: forward: 5'-GAATGGGAAAAGAGGGGAAA-3', reverse: 5'-CC AGCTTCTGGACTCAGCTT-3'; and (3) 558 bp for wild type, undetectable for knockout: forward: 5'-CCCCTTCTTCACTCCATGA-3', reverse: 5'-CTTGG GGACATCTCTTTGAA-3'. The forward primer of (1) and reverse primer of (3) were also combined as the fourth primer pair. The confirmed P19 *Pitx2* knockout cell line was used in this study.

Cytoplasmic and nuclear fraction. In brief, P19 cells were culture in 10-cm plates at 80% confluence. The cells were treated with vehicle (water) or H₂O₂ (300 μM) for 8 h before being collected for cell fraction assay. The NE-PER Nuclear and Cytoplasmic Extraction Reagents (Thermo Scientific) were used according to the manufacturer's manual.

Histology and immunofluorescence. Trichrome staining was performed as previously described²⁸. Fixation, tissue processing, antigen retrieval and blocking for nonspecific staining have been described previously²⁵. Samples were incubated in primary antibody at 4°C overnight. After washing in PBS with Tween 20, sections were incubated in the appropriate fluorescent-labelled secondary antibodies, followed by counterstaining with DAPI (10 μg ml⁻¹) (Roche), and then mounted in VECTASHIELD hardset mounting medium (Vector Laboratories). P19 cells were fixed in formalin (VWR International) for 10 min, then permeabilized in

0.2% Triton X-100 (Bio-Rad Laboratories) in PBS. After blocking in 10% sheep serum (Sigma-Aldrich) for 30 min, cells were incubated with primary antibody for 2 h at room temperature, followed by a 1-h incubation in proper fluorescent-labelled secondary antibodies. Cells were counterstained with DAPI (Roche) then mounted in VECTASHIELD hardest mounting medium (Vector Laboratories). Primary antibodies used were as follows: mouse monoclonal anti- α -Troponin T (1:200; Thermo Scientific), rabbit polyclonal anti-Pitx2 (1:400; Capra Science) and rabbit polyclonal anti-Nrf2 (1:200; Abcam). Secondary antibodies used were as follows: Alexa Fluor 488 goat anti-rabbit IgG and Alexa Fluor 546 donkey anti-mouse IgG (1:400–1:800; Life Technologies); biotinylated anti-mouse IgG (1:200; Vector Laboratories); streptavidin-Alexa Fluor 647 (1:200; Life Technologies). Immunofluorescent images were captured on (1) a Leica TCS SP5 confocal microscope (all functions controlled via Leica LAS AF software (Leica Microsystems)); (2) a Zeiss LSM 510 META laser scanning confocal microscope (all functions controlled via Zeiss LSM Image Browser software (Carl Zeiss Microimaging)); or (3) a Nikon Eclipse 80i upright microscope (all functions controlled by the NIS-Elements BR3.1 software program (Nikon Instruments)). All manuscript figures were prepared using Adobe Photoshop CS5 (Adobe Systems Inc.).

ROS detection. *Pitx2* CKO and control (*Pitx2*^{flg}) 4 DPR hearts were cryo-embedded, and 10 μ m tissue sections were prepared. CellROX green reagent (Life Technologies) was used to detect the presence of ROS according to the manufacturer's manual with minor modifications. Tissue slides were warmed to room temperature (25°C), and rinsed with PBS three times. CellROX substrate was added and incubated for 10 min at 37°C. Slides were given three 5-min washes with PBS, and 10% neutral buffered formalin was added. After a 15 min fixation, mouse-anti-MF20 IgG (1:50; Developmental Studies Hybridoma Bank) and Alexa Fluor 546 donkey anti-mouse IgG (1:400; Life Technologies) were used to counterstain cardiomyocytes. Nuclei were highlighted by DAPI. Slides were mounted in VECTASHIELD hardest mounting medium (Vector Laboratories), and imaged using a Nikon Eclipse 80i upright microscope (Nikon Instruments) within 2 h of completion.

Co-immunoprecipitation. *Pitx2*^{flg} mice were subjected to P2 LAD-O, and ventricular tissue was collected at 4 DPML. P19 cell fractions were obtained as described above. For Flag pull-down, anti-Flag M2 affinity gel (Sigma-Aldrich) was used. For Nrf2 pull-down, rabbit polyclonal anti-Nrf2 (Abcam) and protein A/G PLUS-Agarose (Santa Cruz Biotechnology) were used according to the manufacturer's manual.

Western blot. Ventricles of 6 DPR ($n = 3$) and sham ($n = 3$) *Pitx2*^{flg} mice, P16 *Mhc*^{cre-Ert} (control, $n = 3$) and *Pitx2*-overexpressing ($n = 3$), as well as P19 cells were collected and lysed in RIPA buffer, and the protein concentration was quantified using Pierce BCA protein assay kit (Pierce Biotechnology) as previously described²⁵, the P19 cell fraction were acquired as described earlier, with three biological replicates. Co-immunoprecipitation samples were acquired as described above. In brief, after separation via SDS-PAGE, proteins were transferred to PVDF membranes (EMD Millipore), blocked in 5% milk/TBS-Tween 20 and incubated with appropriate primary antibodies (all with 1:1,000 dilution in TBST) overnight at 4°C (rabbit anti-Flag IgG and mouse anti- α -tubulin IgG, Sigma-Aldrich; rabbit anti-Pitx2 IgG, Capra Science; rabbit anti-Yap, Novus Biologicals; rabbit anti-TATA-binding protein, Cell Signaling Technology; rabbit polyclonal anti-Nrf2, Abcam). Membranes were then washed three times in TBST and incubated with goat-anti-rabbit and goat-anti-mouse horseradish peroxidase (HRP)-conjugated secondary antibodies (1:5,000; Santa Cruz Biotechnology) for 1 h at room temperature. Protein detection was performed using SuperSignal West Pico Chemiluminescent Substrate (Thermo Scientific). For quantification, Pitx2 and α -tubulin band densities for control (*Mhc*^{cre-Ert}) and *Pitx2*-overexpressing mice were determined using ImageJ software (National Institutes of Health).

ChIP. P1 apex resection or sham surgery was performed on *Pitx2*^{flg} neonates; P2 LAD-O was performed on C57BL6 neonates. At 5 days after surgery, whole ventricles were collected and subjected to ChIP assay using the EZ-ChIP kit (Millipore) according to the manufacturer's protocol. For Flag ChIP, anti-Flag M2 affinity gel (Sigma-Aldrich) was used; for Nrf2 ChIP, rabbit polyclonal anti-Nrf2 antibody (Abcam) was used; for Yap ChIP, rabbit polyclonal anti-Yap antibody (Novus Biologicals) was used.

ChIP-re-ChIP. P2 LAD-O was performed on *Pitx2*^{flg} neonates, and 4 days later whole ventricles were collected and subjected to the ChIP-re-ChIP assay as previously described²⁵. Monoclonal anti-Flag BioM2 antibody (Sigma-Aldrich) was used for pulling down Flag; rabbit anti-YAP (Novus Biologicals) was used for pulling down Yap. For qPCR analysis of the peak and non-peak regions, the following primers were used. For peak region: *Oxnd1*, forward: 5'-GGGTTTTAGTGGGCAACCTAT-3', reverse: 5'-CTGGGCTTTAGA GACAGCTAGG-3'; *Ldha*, forward: 5'-CCAGAGATCTTGTCCAGTCCCTT-3', reverse: 5'-TTCAGTTCCAAATGGGGATAC-3'; *Ndufb3*, forward: 5'-GTACCGG AACGTTTACCATCTC-3', reverse: 5'-CACCAGCTCCCTAAATTACCTG-3';

Ndufs8, forward: 5'-CATAGTGCCTTTCTCTTCTTG-3', reverse: 5'-GGCGCATTAACCTCTCTGATAC-3'.

For non-peak-region: *Oxnd1*, forward: 5'-TAATGAGATCTGGTGCCCT CTT-3', reverse: 5'-GACACACCCCATCTGTACTTCA-3'; *Ldha*, forward: 5'-GTATCTTCACAGGCCCTTCCTG-3', reverse: 5'-GGCTGTGGGACAAATGTT CTTAT-3'; *Ndufb3*, forward: 5'-TGCTACTCTTCCAGAGGACCTT-3', reverse: 5'-GGTATGTGTGTGTGTGATGTGC-3'; *Ndufs8*, forward: 5'-TCTACTGCCTTT ATGGCGTTT-3', reverse: 5'-CTCATGGGCTGTGACAAATAGAA-3'.

qPCR. For Fig. 1h, i, mRNA was prepared from P19 cells, and control (C57BL6) and *Nrf2*^{nu/nu} ventricles at P1. For Fig. 4d, DNA was prepared as described in the 'ChIP-re-ChIP' section. For Supplementary Fig. 9e, DNA was prepared according to the ChIP protocol as mentioned above. For Supplementary Fig. 2b, mRNA of *Mhc*^{cre-Ert} (control) and *Pitx2*-overexpressing ventricles was prepared at P16, 8 days after the first tamoxifen injection (daily from P7–P10). For Supplementary Fig. 9f, g, mRNA was prepared from P19 cells and ES cells. Total mRNA was extracted using miRNeasy Mini Kit (Qiagen); cDNA was generated using qScript cDNA supermix (Quanta BioSciences); qPCR was performed on a StepOnePlus Real-Time PCR system (Applied Biosystems) with iTaq Universal SYBR Green Supermix (Bio-Rad Laboratories), all according to the manufacturers' manuals. The primers used for ChIP-qPCR are as followed: *Gpx1* region1, forward: 5'-GCTTCATCCCTCCTAATGGA-3', reverse: 5'-TGCCAGCATTAACCTCAGAGC-3'; *Gpx1* region2, forward: 5'-TCTTCTTAGG CGGGACTCTA-3', reverse: 5'-GGGTCTGGTCTAGCTCCTGT-3'; *Mt1*, forward: 5'-TTCTGTCAGTCCAGTCTGACC-3', reverse: 5'-ATAGGAGATGGCC TGGTGAC-3'; *Mt2* region1, forward: 5'-GCCCTCCCACCTACTCATTA-3', reverse: 5'-GGTGACTGTATCCCACTTG-3'; *Mt2* region2, forward: 5'-TTCACT AAGAGTGCAGGA-3', reverse: 5'-ATCTGCAGAGCCAGGAAACT-3'; *Sod2* region1, forward: 5'-ACGTGGCTTCAGGAGATTT-3', reverse: 5'-CAATAT CGCTTGCTCTCAGC-3'; *Sod2* region2, forward: 5'-CTCTCATGCATGCA AATCCT-3', reverse: 5'-CAGCTCTAAGGGACCCAGAC-3'; *Sod1* region1, forward: 5'-ACTGTGACCTTGCAAAAACA-3', reverse: 5'-GTCCACCACTTC AGAGAGCA-3'; *Txn2* region1, forward: 5'-CCACACAGCTGAAGGAGAGA-3', reverse: 5'-GGAGTGTGGGAATGTAGGT-3'; *Txn2* region2, forward: 5'-CCAAAATACCCAAGCCTGTT-3', reverse: 5'-TTTCCACATGCCTCTGTG TC-3'; *Ndufv1* region1, forward: 5'-GGCTGCGAGGAAGAAATAAC-3', reverse: 5'-ACTAACGGTCCCAATCCAG-3'; *Ndufv1* region2, forward: 5'-ACAAGATGCAGGTTCATGGAA-3', reverse: 5'-ATCAGAGCCACACTG TCTGC-3'; *Cox5a*, forward: 5'-GCTGTTCTGGGATTGGATCT-3', reverse: 5'-AGAGCCTGTCTCTCCAAAA-3'.

The primers used for other qPCR are as followed: *Gpx1*, forward: 5'-GTCCACCGTGTATGCCTTCT-3', reverse: 5'-CTCCTGGTGTCCGAACGTG AT-3'; *Mt2*, forward: 5'-CCGATCTCTCGTCGATCTC-3', reverse: 5'-AGGAGCAGCAGCTTTTCTTG-3'; *Mt1*, forward: 5'-GCTGTCTCTAAGC GTCACC-3', reverse: 5'-AGGAGCAGCAGCTCTTCTTG-3'; *Sod2*, forward: 5'-GGCCAAGGAGATGTTACAA-3', reverse: 5'-GCTTGATAGCCTCCA GCAAC-3'; *Txn2*, forward: 5'-CCCCTCAGTACAATGCTGGT-3', reverse: 5'-TCCATCTGGACGTTAAAG-3'; *Pitx2*, forward: 5'-AGGGAGGGAGG CAAGAAAAG-3', reverse: 5'-CTTGAAAGAGCCAGGGAACG-3'; *Nrf2*, forward: 5'-CCAGAAGCCACACTGACAGA-3', reverse: 5'-GGAGAGGATG CTGCTGAAAG-3'.

RNA-seq and ChIP-seq. Total mRNA was extracted from the ventricles of *Pitx2* CKO and *Pitx2*^{flg} mice 5 days after P1 apex resection, using the miRNeasy Mini kit (Qiagen); ChIP DNA was acquired as described above. RNA-seq and ChIP-seq were performed using the Ion Proton system for next-generation sequencing according to the manufacturer's direction. Sequenced reads were mapped to mm9 genome using Ion Torrent TMAP aligner with 'map4' option. We used HTSeq-Count (version 0.5.4) to quantify the aligned RNA-seq reads against exon regions of genes in RefSeq mm9 annotation. Differential expressed genes were detected using R package DESeq with threshold $P \leq 0.05$, fold change ≥ 1.5 and FDR $\leq 10\%$. ChIP-seq peaks were detected by Homer package 'findPeak' command using threshold FDR $\leq 10\%$. Only peaks detected from both biological replicates were annotated and overlaid with differential gene expression list. GO analysis was performed on DAVID online platform. Terms with $P \leq 0.05$ were included. Published data sets used in Fig. 1f were obtained from Gene Expression Omnibus (GEO) (series GSE52386, reviewed in ref. 32). Mapped human DHS-seq data were extracted from the GEO (GSE18927, GSE32970). In total, 1,002 genes were downregulated in *Pitx2* CKO 5 DPML and were overlaid with *Pitx2* ChIP-seq binding genes, which were further overlaid with Yap ChIP-seq from control 5 DPML. We define genes that were co-regulated by both *Pitx2* and Yap as expression level decreased in *Pitx2* CKO heart and bound by both factors on the promoter regions. Overrepresented GO analysis was performed using online tool DAVID 6.7 (<https://david.ncifcrf.gov/>).

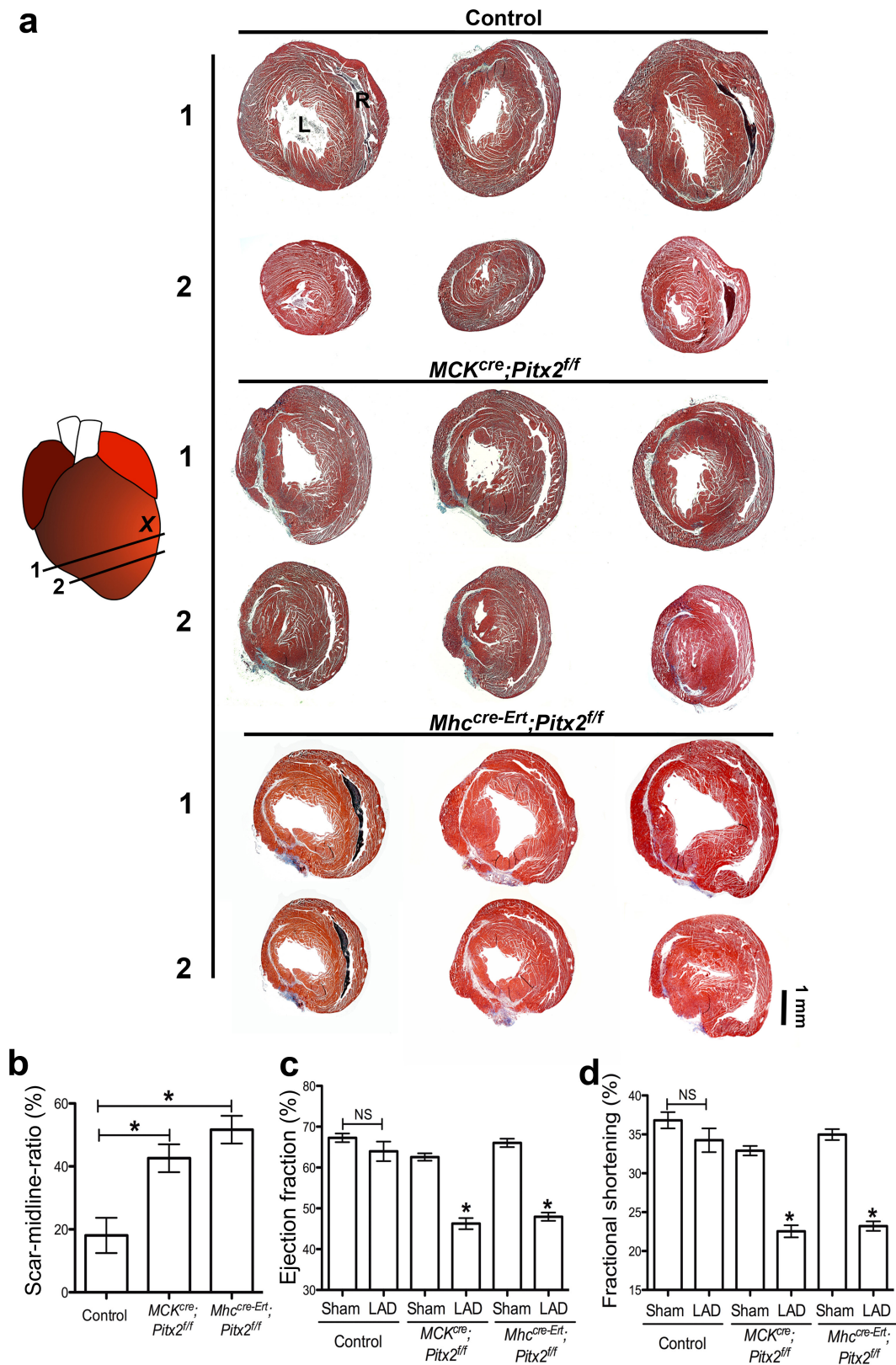
GST pull-down assay. The mouse Yap, Pitx2a, Pitx2c and truncated proteins were prepared as previously described³³. In brief, GST-tagged proteins were expressed in and purified from BL21 competent *Escherichia coli* (New England Biolabs), using Glutathione Sepharose 4B (GE Healthcare). YAP was cleaved from GST using PreScission Protease (GE Healthcare), and 1 µg was incubated with 15 µg of each truncated GST-Pitx2 protein for the pull-down assay. Incubation of corresponding truncated Pitx2 protein alone was used as controls. Purified YAP protein (2 µg) was loaded into the gel as a control. Rabbit-anti-YAP antibody (Cell Signaling Technology) was used for immunoblotting and the detection of YAP.

Coomassie blue staining. GST-Pitx2, GST-Yap and cleaved Yap protein were run on 10% SDS-PAGE gel. The gel was then stained for protein in Coomassie blue stain (2.5 g l⁻¹ in H₂O:methanol:glacial acetic acid at a ratio of 9:9:2) for 1 h with shaking, followed by destaining with Coomassie solvent (H₂O:methanol:glacial acetic acid = 9:9:2) for 2 h with shaking. The stained gel was scanned with EPSON Perfection 4490 Photo (Epson America).

Statistics. Each experimental group in the ChIP-seq and RNA-seq studies had $n = 2$. All quantitative experiments (for example, qPCR, western blot, cell count) have at least three independent biological repeats. For animal studies (neonatal and adult surgery), sample sizes were estimated based on our pilot studies. The n number for each experiment is summarized in relevant sections in the Methods. Differences between groups were examined for statistical significance using non-parametric test (Mann-Whitney test) (for two groups), or one-way ANOVA plus Bonferroni post-test (for more than two groups). Equal variances were assumed (no Welch's correction). Grubbs's test was used to determine outlier (GraphPad Prism, GraphPad Software). In the case of Fig. 4d, Wilcoxon signed-rank test (with

a hypothetical value of 0) was used to compare anti-Yap plus anti-Flag group to anti-IgG group, as well as anti-Yap-anti-IgG group, since the latter two groups were undetected in qPCR assay, the same test (Wilcoxon signed-rank) was also applied to Fig. 1h, i (left panel) (with a hypothetical value of 1). All bar graphs represent mean \pm s.e.m. * $P < 0.05$, *** $P < 0.001$ were considered statistically significant.

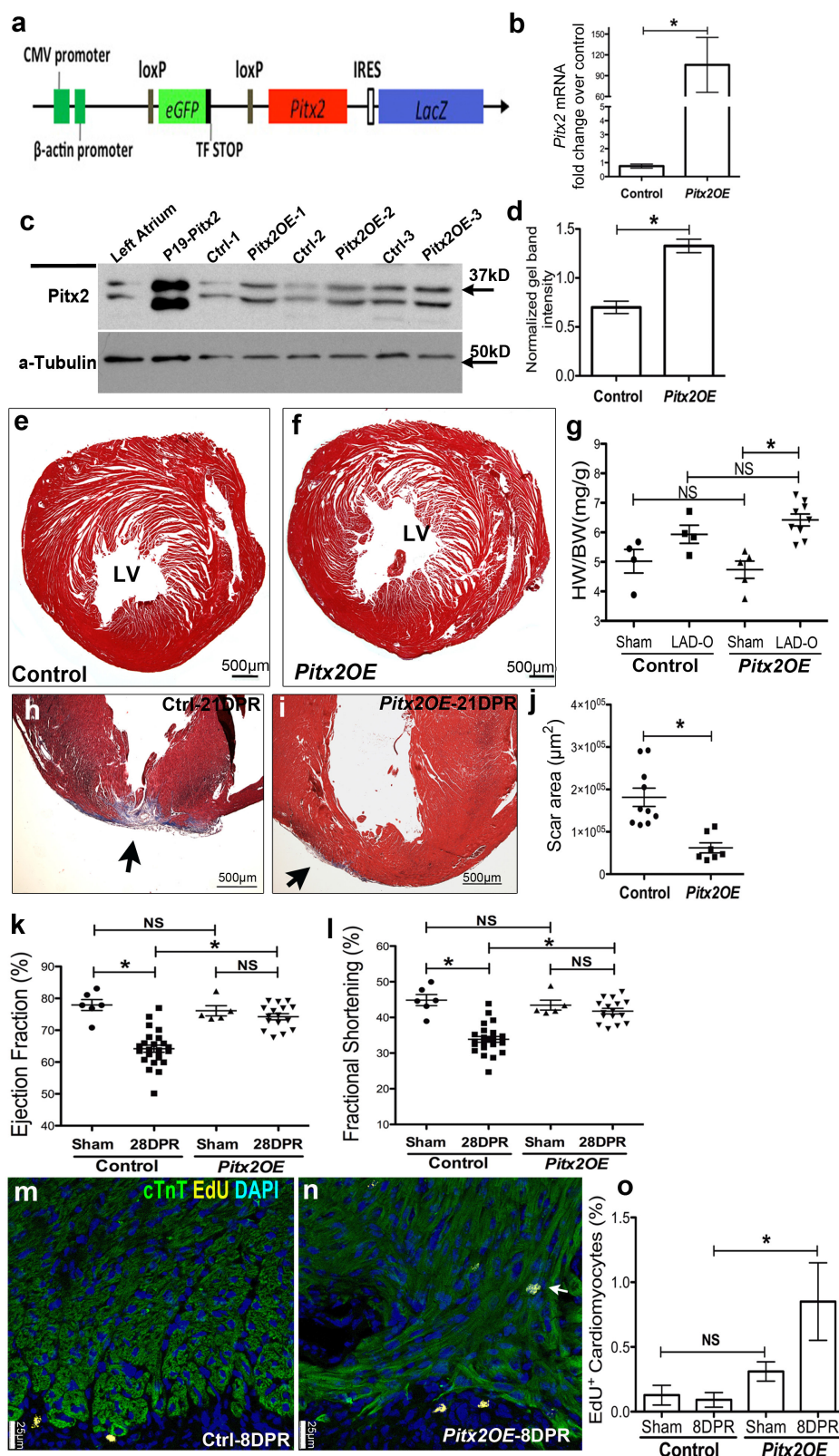
25. Heallen, T. *et al.* Hippo pathway inhibits Wnt signaling to restrain cardiomyocyte proliferation and heart size. *Science* **332**, 458–461 (2011).
26. Lavado, A., Lagutin, O. V., Chow, L. M., Baker, S. J. & Oliver, G. Prox1 is required for granule cell maturation and intermediate progenitor maintenance during brain neurogenesis. *PLoS Biol.* **8**, e1000460 (2010).
27. Xin, M. *et al.* Regulation of insulin-like growth factor signaling by Yap governs cardiomyocyte proliferation and embryonic heart size. *Sci. Signal* **4**, ra70 (2011).
28. Heallen, T. *et al.* Hippo signaling impedes adult heart regeneration. *Development* **140**, 4683–4690 (2013).
29. Nascimento, D. S. *et al.* MIQuant—semi-automation of infarct size assessment in models of cardiac ischemic injury. *PLoS ONE* **6**, e25045 (2011).
30. Porrello, E. R. *et al.* Regulation of neonatal and adult mammalian heart regeneration by the miR-15 family. *Proc. Natl Acad. Sci. USA* **110**, 187–192 (2013).
31. Ran, F. A. *et al.* Genome engineering using the CRISPR-Cas9 system. *Nature Protocols* **8**, 2281–2308 (2013).
32. Romanoski, C. E., Glass, C. K., Stunnenberg, H. G., Wilson, L. & Almouzni, G. Epigenomics: Roadmap for regulation. *Nature* **518**, 314–316 (2015).
33. Amen, M. *et al.* Chromatin-associated HMG-17 is a major regulator of homeodomain transcription factor activity modulated by Wnt/ β -catenin signaling. *Nucleic Acids Res.* **36**, 462–476 (2008).



Extended Data Figure 1 | *Pitx2* is required in neonatal myocardial regeneration after LAD-O. **a**, Serial trichrome images of control (*Pitx2^{fl/fl}*), *MCK^{cre};Pitx2^{fl/fl}*, and *Mhc^{cre-Ert};Pitx2^{fl/fl}* 21 days after LAD-O performed in P2 mice. Three representative hearts of each genotype were shown.

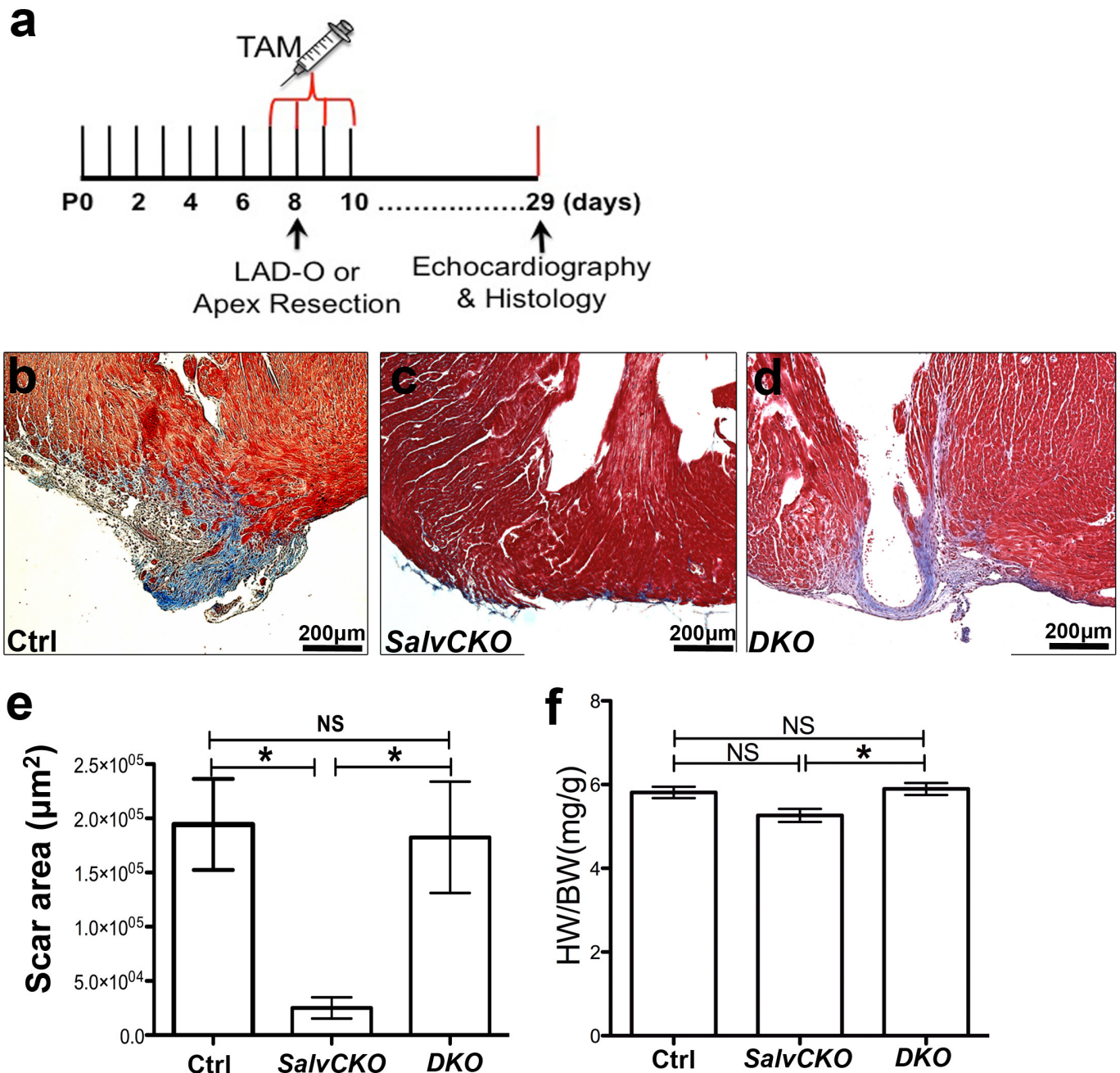
b, Percentage of fibrotic left ventricular myocardium quantified at 3 weeks

after LAD-O; $n = 8$ for control (*Pitx2^{fl/fl}*), $n = 7$ for *MCK^{cre};Pitx2^{fl/fl}*, and $n = 4$ for *Mhc^{cre-Ert};Pitx2^{fl/fl}*. **c**, **d**, Ejection fraction (**c**) and fractional shortening (**d**) of LAD-O and sham hearts (see Methods for n). L, left ventricle; R, right ventricle. Mean \pm s.e.m. * $P < 0.05$ one-way ANOVA plus Bonferroni post-test (**c**, **d**) and Mann-Whitney test (**b**).



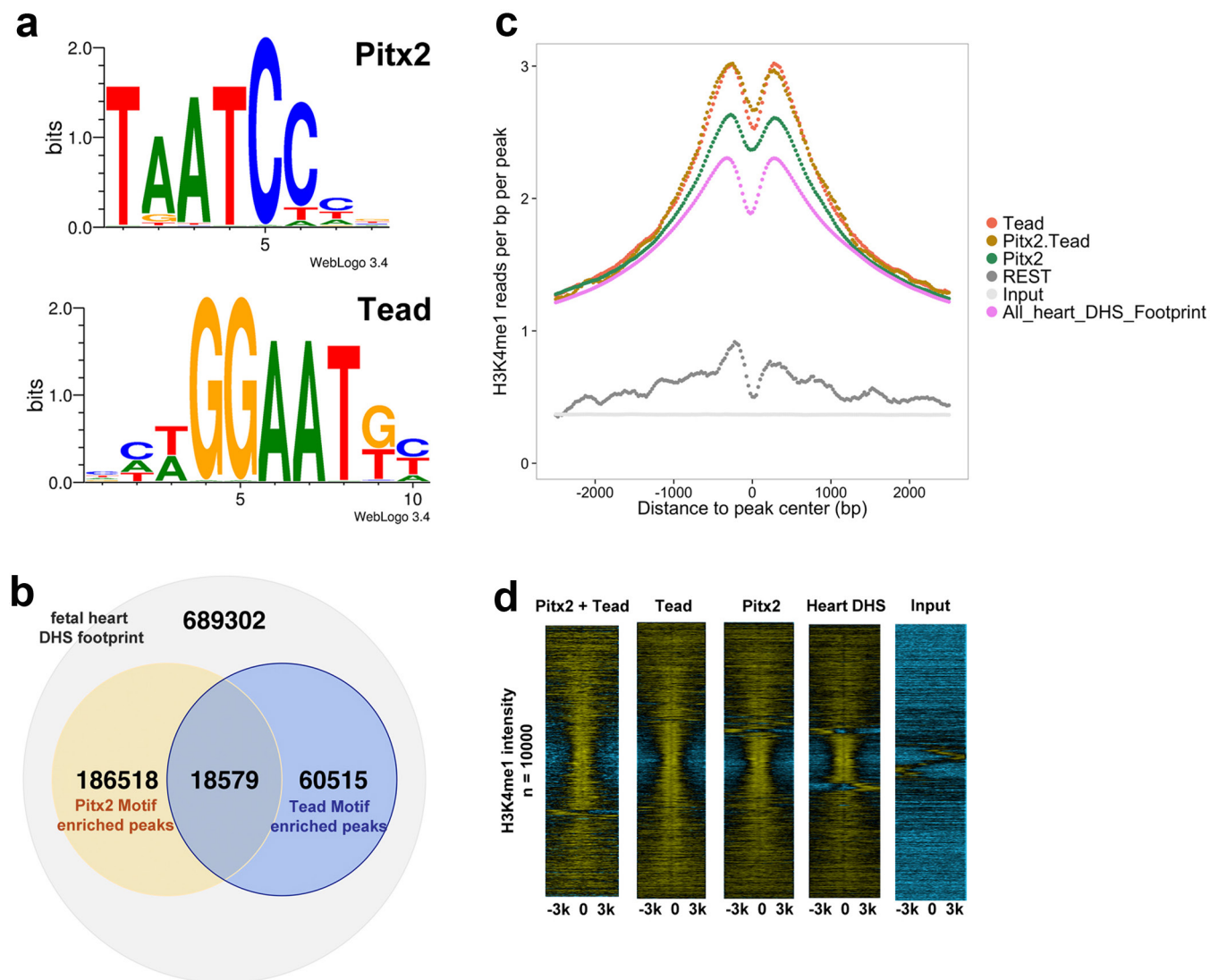
Extended Data Figure 2 | *Pitx2* promotes myocardial regeneration after apex resection at P8. **a**, Schematic of *Pitx2*-expressing construct (*Pitx2*^{Gof}). **b–d**, *Pitx2*^{Gof} was crossed with the *Mhc*^{cre-Ert} strain to generate *Mhc*^{cre-Ert/+};*Pitx2*^{Gof} (*Pitx2*-overexpressing) mice. After tamoxifen treatment from P7–P10, qPCR (**b**, *n* = 4) and western blot (**c**, **d**, *n* = 3) show the overexpression of *Pitx2* in the myocardium at P16. **e**, **f**, Trichrome-stained cross sections from 13-week-old sham hearts of control (**e**) and *Pitx2*-overexpressing (**f**) mice, with tamoxifen administrated at 7–8 weeks old. **g**, Heart weight over body weight ratio of adult sham and LAD-O hearts; *n* = 4 (control sham), 4 (control LAD-O), 5 (*Pitx2*-overexpressing sham),

9 (*Pitx2*-overexpressing LAD-O). **h–j**, Apex resection of *Pitx2*-overexpressing (**i**) and control (*Mhc*^{cre-Ert/+}) (**h**) hearts at P8 followed by trichrome staining at 28 DPR; the scar area was quantified in **j**; *n* = 10 (control mice), 7 (*Pitx2*-overexpressing mice). **k**, **l**, Echocardiography showed ejection fraction (**k**) and fractional shortening (**l**) at 28 DPR (see Methods for *n*). **m–o**, EdU labelling of *Pitx2*-overexpressing (**n**) and control (**m**) apical area, 8 days after P8 resection, sections were stained for cTnT (green), EdU (yellow), and DAPI (blue). Arrow indicates EdU-labelled cardiomyocytes, with quantification in **o**; *n* = 4 mice per group. Mean \pm s.e.m. **P* < 0.05, one-way ANOVA plus Bonferroni post-test (**g**, **k**, **l**) and Mann–Whitney test (**b**, **d**, **j**, **o**).



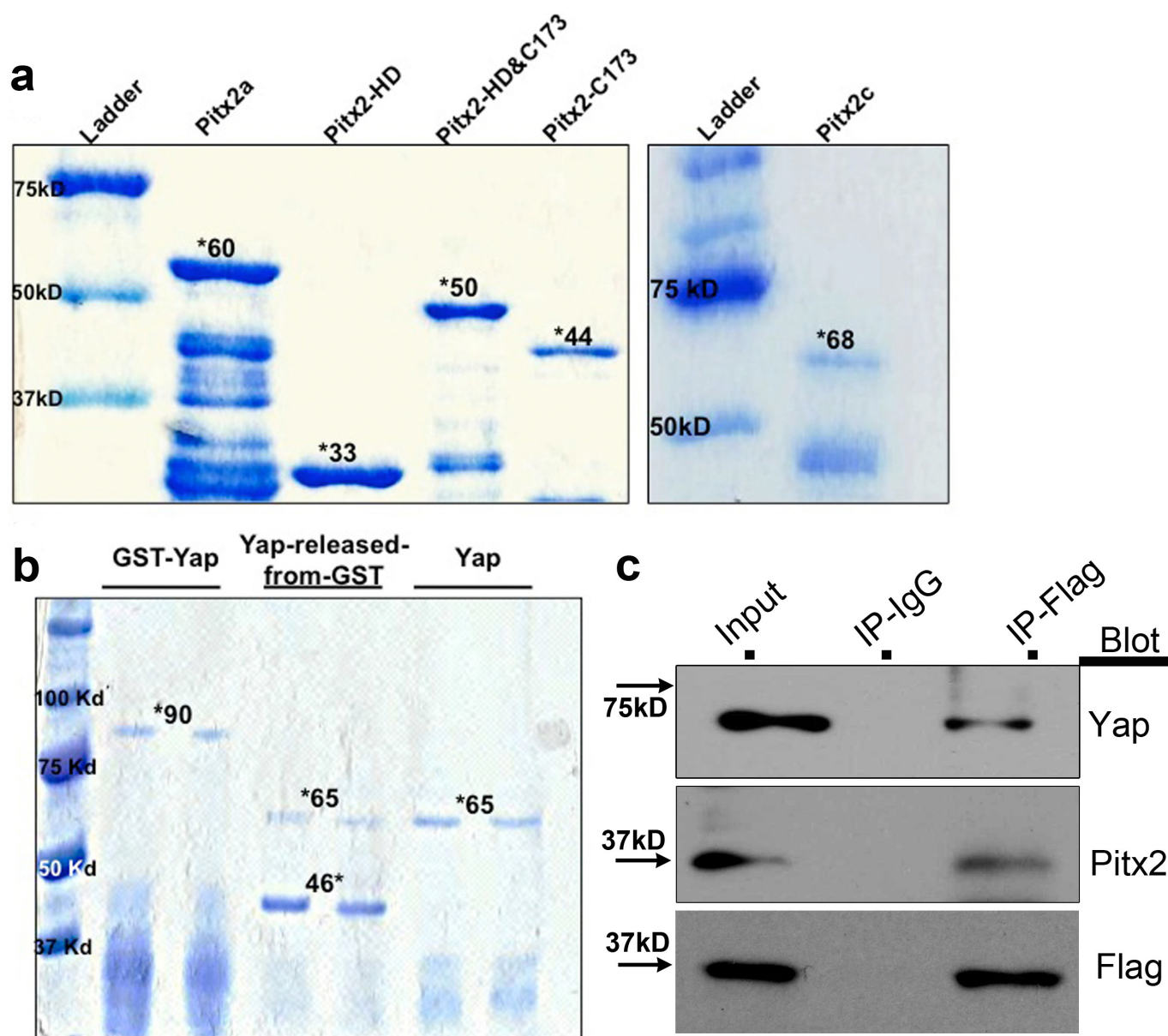
Extended Data Figure 3 | Pitx2 is required for Hippo-deficient heart regeneration. **a**, Schematic study plan for Fig. 3a–e. **b–e**, Trichrome-stained apical areas of control (**b**), *Salv* CKO (**c**) and double knockout (**d**) hearts 21 days after P8 apex resection. Scar area was quantified in

e, f, Heart weight to body weight ratio of sham hearts at 28 days after tamoxifen administration. For *n* number, see Methods. Mean ± s.e.m. **P* < 0.05, Mann–Whitney.



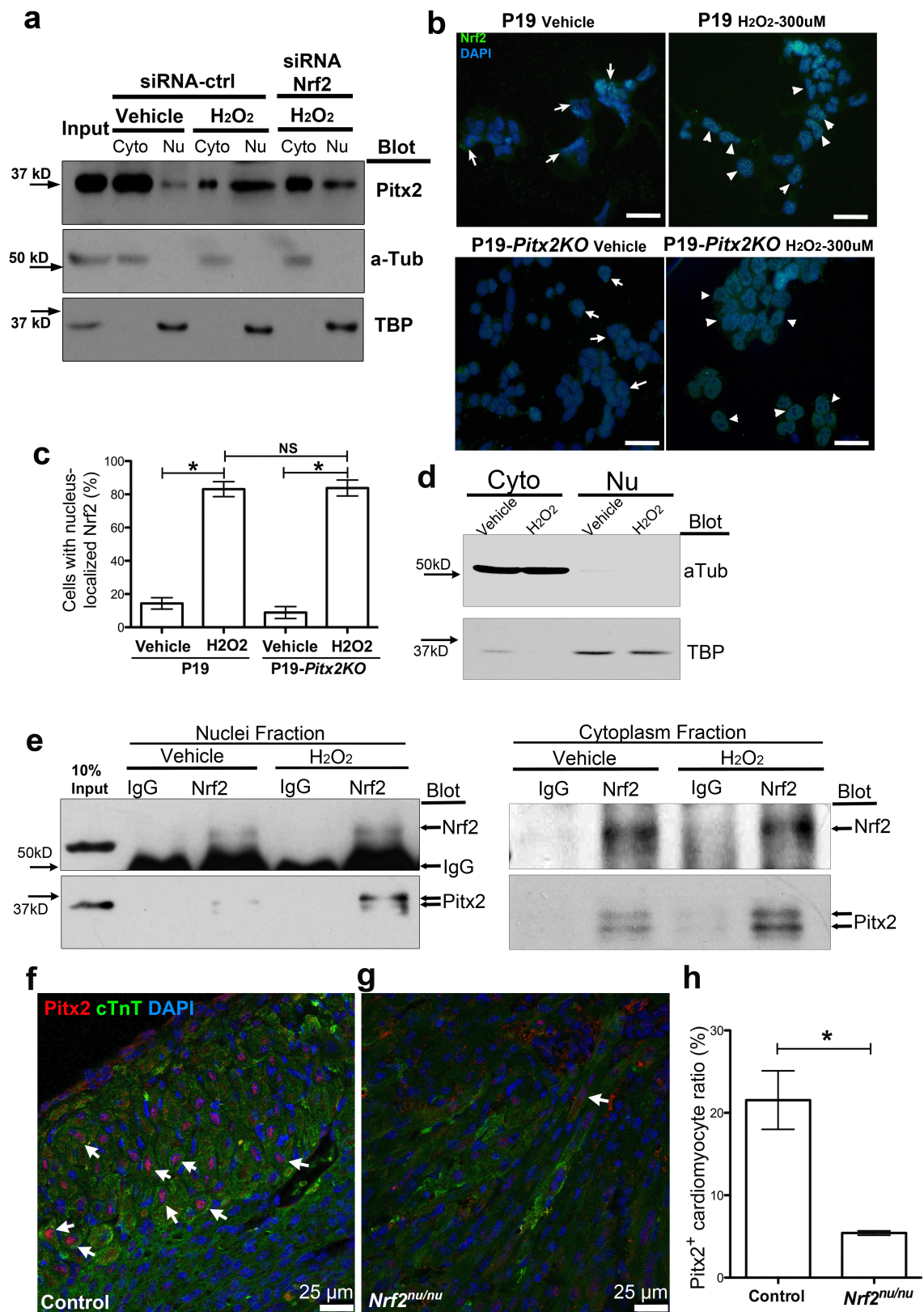
Extended Data Figure 4 | Co-occurrence of Pitx2 and Tead DNA-binding motifs in fetal heart enhancers. **a**, Consensus *Pitx2* and *Tead* motifs. **b**, *Pitx2* and *Tead* motif co-occurrence in fetal heart DHS peaks. **c**, Aggregate plot of H3K4me1 in fetal heart ChIP-seq reads within 6 kb range of DHS peaks. **d**, Heat map of fetal heart H3K4me1 ChIP-seq or

input read density in 6-kb regions of DHS peaks. DHS peaks were centred on the *Pitx2* motif, *Tead* motif, *Pitx2-Tead* motifs, or randomly selected. The read density was in \log_2 scale. Blue, negative values; yellow, positive values.



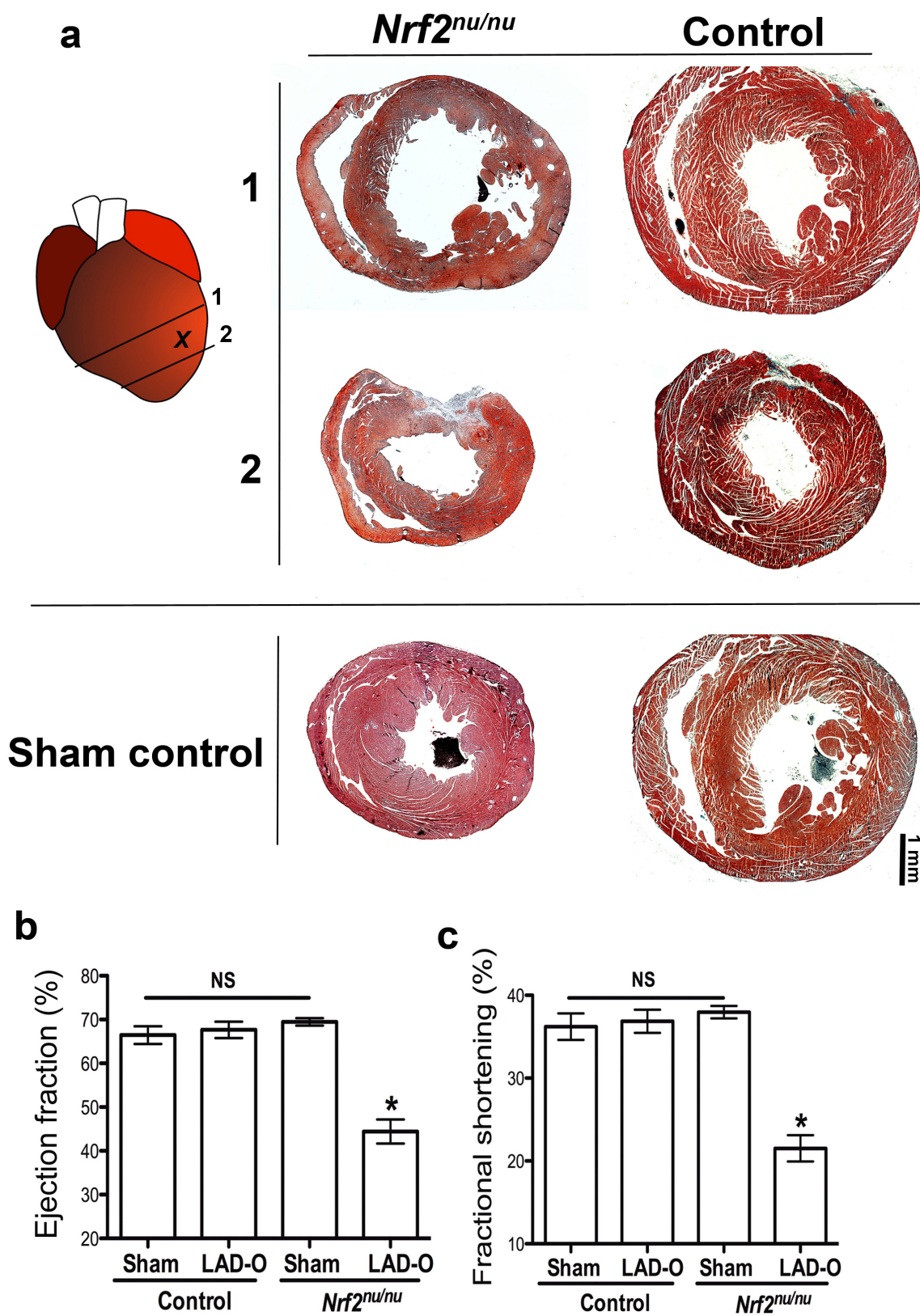
Extended Data Figure 5 | Generation of GST-tagged proteins and interaction between Pitx2 and Yap *in vivo*. **a**, The mouse Pitx2a, Pitx2c and truncated proteins were purified and run on a 10% SDS-PAGE gel, and Coomassie blue staining shows the GST fusion protein band with

correct size (marked by asterisk). **b**, Coomassie blue staining of the purified GST-Yap, Yap cut by prescission protease and pure Yap protein. **c**, Co-immunoprecipitation of Flag in *Pitx2^{flag}* ventricles at 5 DPR, and blotting of Yap, Pitx2 and Flag.

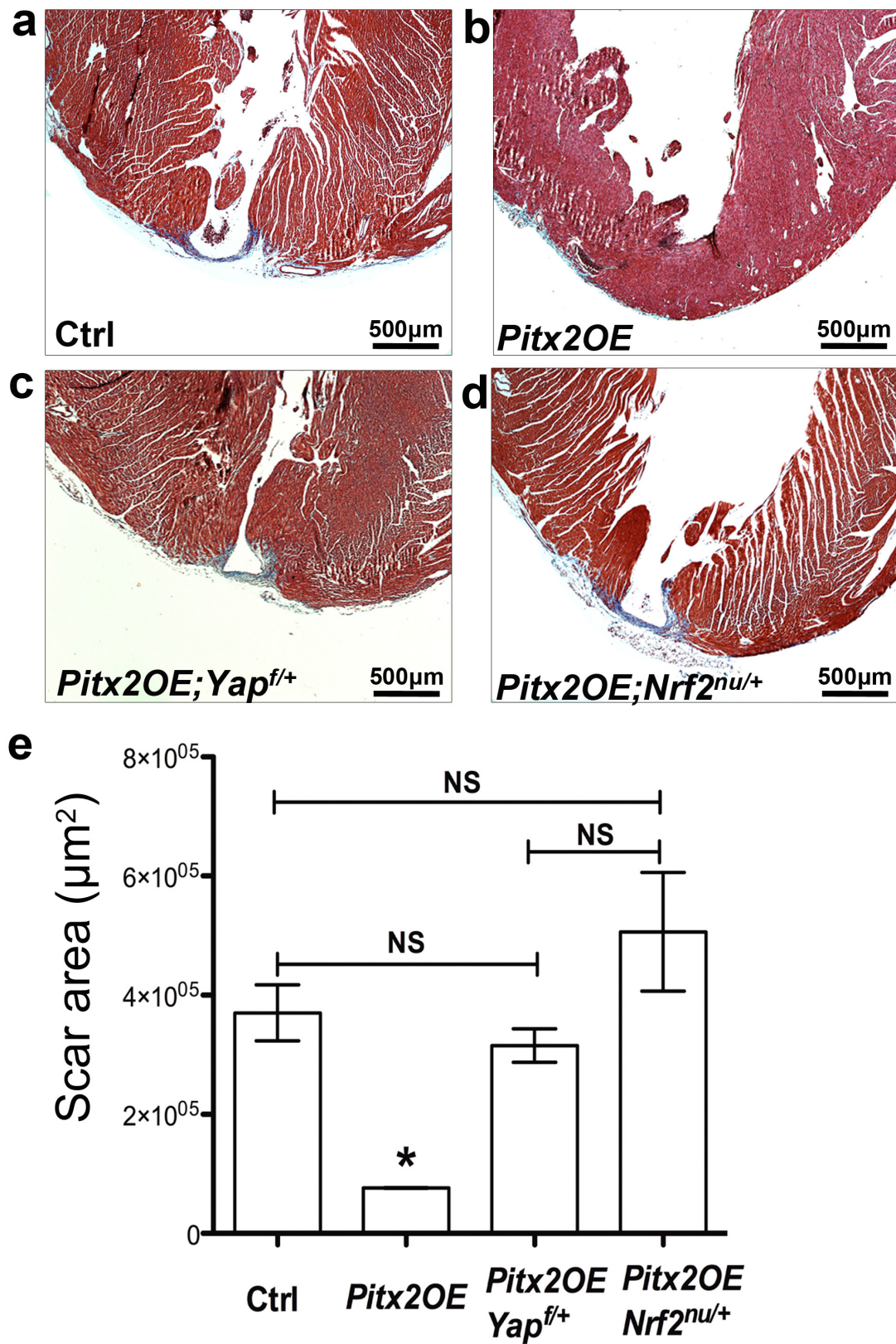


Extended Data Figure 6 | Nuclei-shuttling of Nrf2 is independent of *Pitx2*. **a**, Western blotting of Pitx2, α -tubulin, and TATA-binding protein (TBP) of P19 cell fraction after H₂O₂, with or without Nrf2 siRNA treatment. **b**, Immunofluorescent staining of Nrf2 (green) in P19 control and *Pitx2* knockout cells after vehicle or H₂O₂ treatment. DAPI, blue. Scale bars, 50 μ m. **c**, The ratio of cells with nuclear Nrf2 over total cell number; $n = 6$ biological repeats. **d**, Blotting of α -tubulin and TBP to show cell fraction of P19 cells used in **e**. **e**, Co-immunoprecipitation

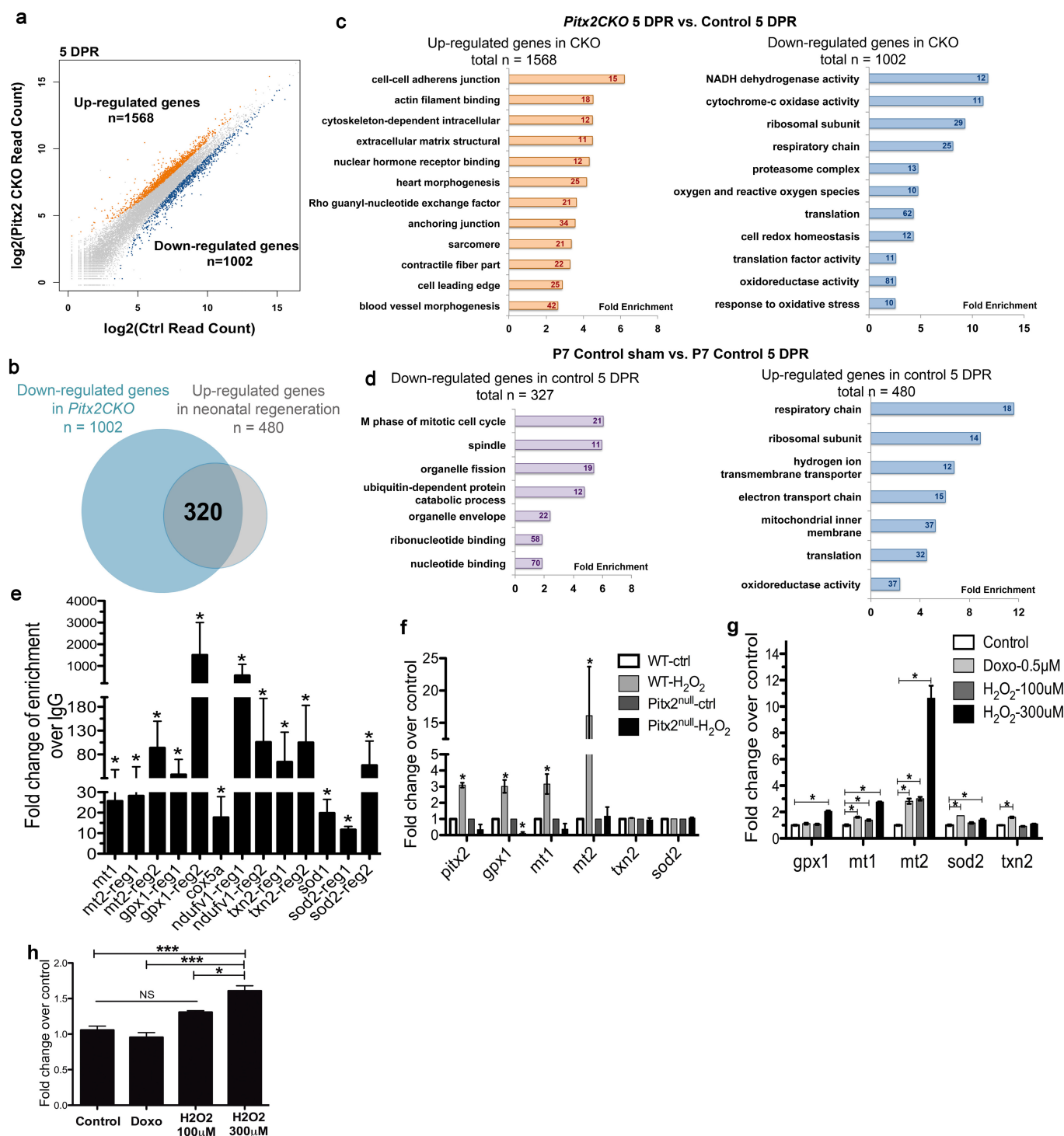
Nrf2 from nuclear and cytoplasmic fraction of P19 cells after vehicle or H₂O₂ treatment, blotting shows Nrf2 and Pitx2. **f–h**, 4 DPMI control (C57BL6) (**f**) and *Nrf2*^{nu/nu} (**g**) cross-sections stained for Pitx2 (red), cTnT (green), and DAPI (blue), with the ratio of cardiomyocytes with nuclei-localized Pitx2 quantified in **h**; $n = 4$ mice per group. Arrows, Pitx2⁺ cardiomyocytes. Mean \pm s.e.m. * $P < 0.05$, one-way ANOVA plus Bonferroni post-test (**c**) and Mann–Whitney test (**h**).



Extended Data Figure 7 | Nrf2 is required for neonatal myocardial regeneration. **a**, Trichrome images of *Nrf2^{nu/nu}* and control heart (C57BL6) at 21 days after P2 LAD-O, along with sham controls. **b**, **c**, Ejection fraction (**b**) and fractional shortening (**c**) of LAD-O and sham hearts (see Methods for *n*). Mean \pm s.e.m. * $P < 0.05$, Mann-Whitney test.



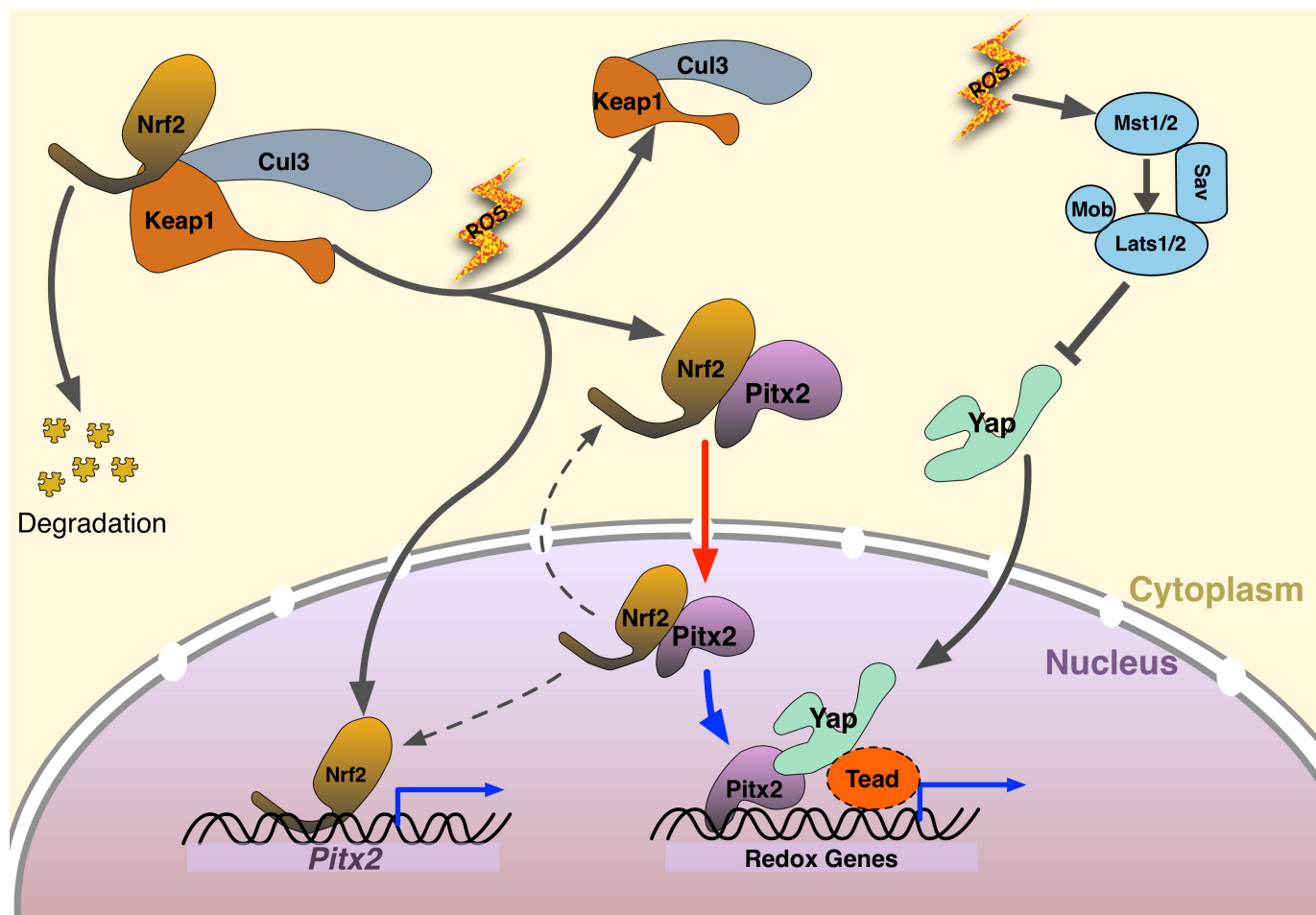
Extended Data Figure 8 | *Yap1* and *Nrf2* are essential for *Pitx2*-induced myocardial regeneration. a–d, Trichrome staining showing apical scarring of different groups at 28 DPR, apex resection was performed at P8. e, Quantification of scar area; $n = 4$ mice per group. Mean \pm s.e.m. * $P < 0.05$, Mann–Whitney test, control compared to the other three groups individually.



Extended Data Figure 9 | *Pitx2* regulates antioxidant scavenger genes.

a, Overall change of genes in *Pitx2* CKO mice compared to control.
b, Upregulated genes in 5 DPR control over wild-type sham heart ($n = 480$) overlaid with downregulated genes in 5 DPR *Pitx2* CKO over 5 DPR control heart ($n = 1,002$). **c**, GO analysis of genes upregulated (left) and downregulated (right) in *Pitx2* CKO ventricles over controls at 5 DPR.
d, GO analysis of genes upregulated (right) and downregulated (left) in 5 DPR control ventricles over age matching sham hearts. **e**, ChIP-qPCR

confirming the binding of *Pitx2* to the regulatory regions of target genes; $n = 4$ biological replicates. **f**, qPCR detecting *Pitx2* and antioxidant genes in wild-type and *Pitx2*^{nu/nu} ES cells after vehicle or H₂O₂ treatment; $n = 4$ biological replicates. **g**, qPCR of antioxidant genes in P19 cells after doxorubicin or H₂O₂ treatment; $n = 5$ biological replicates. **h**, qPCR of *Pitx2* in P19 cells after doxorubicin or H₂O₂ treatment; $n = 5$ biological replicates. Mean \pm s.e.m. * $P < 0.05$; *** $P < 0.001$, Mann-Whitney test.



Extended Data Figure 10 | Mechanism model of *Pitx2*, *Nrf2* and *Yap1* responding to oxidative stress. When oxidative stress is low, *Nrf2* is sequestered in cytoplasm by its degradation complex (*Cul3*, *Keap1*), and *Pitx2* stays either in the cytoplasm or at low expression levels. When the redox balance is disturbed by ROS, *Nrf2* breaks away from the degradation complex, and enters nuclei to upregulate *Pitx2* gene expression; *Nrf2* also binds cytoplasmic *Pitx2* and shuttles it to the nuclei, where *Pitx2* and *Yap* co-regulate their common targets including critical antioxidant genes.

In wild-type adult mouse heart, active *Yap* is maintained at a low level, even after ischaemic injury, and is thus not able to repair myocardium efficiently. When *Pitx2* is overexpressed in cardiomyocytes, sufficient amounts of *Pitx2* will cooperate with low levels of resident active *Yap* to induce the expression of beneficial antioxidant scavengers in a synergetic pattern, rendering protection to injured myocardium. Red arrow, supported by *in vitro* evidence; Blue arrows, supported by *in vivo* evidence.

Feedback modulation of cholesterol metabolism by the lipid-responsive non-coding RNA *LeXis*

Tamer Sallam^{1,2*}, Marius C. Jones^{1*}, Thomas Gilliland¹, Li Zhang¹, Xiaohui Wu^{1,2}, Ascia Eskin³, Jaspreet Sandhu¹, David Casero¹, Thomas Q. de Aguiar Vallim², Cynthia Hong¹, Melanie Katz⁴, Richard Lee⁴, Julian Whitelegge⁵ & Peter Tontonoz¹

Liver X receptors (LXR) are transcriptional regulators of cellular and systemic cholesterol homeostasis. Under conditions of excess cholesterol, LXR activation induces the expression of several genes involved in cholesterol efflux¹, facilitates cholesterol esterification by promoting fatty acid synthesis², and inhibits cholesterol uptake by the low-density lipoprotein receptor³. The fact that sterol content is maintained in a narrow range in most cell types and in the organism as a whole suggests that extensive crosstalk between regulatory pathways must exist. However, the molecular mechanisms that integrate LXRs with other lipid metabolic pathways are incompletely understood. Here we show that ligand activation of LXRs in mouse liver not only promotes cholesterol efflux, but also simultaneously inhibits cholesterol biosynthesis. We further identify the long non-coding RNA *LeXis* as a mediator of this effect. Hepatic *LeXis* expression is robustly induced in response to a Western diet (high in fat and cholesterol) or to pharmacological LXR activation. Raising or lowering *LeXis* levels in the liver affects the expression of genes involved in cholesterol biosynthesis and alters the cholesterol levels in the liver and plasma. *LeXis* interacts with and affects the DNA interactions of RALY, a heterogeneous ribonucleoprotein that acts as a transcriptional cofactor for cholesterol biosynthetic genes in the mouse liver. These findings outline a regulatory role for a non-coding RNA in lipid metabolism and advance our understanding of the mechanisms that coordinate sterol homeostasis.

It is well established that the cholesterol biosynthetic pathway is downregulated under conditions in which sterols are abundant through the inhibition of sterol regulatory element-binding protein (SREBP) processing⁴. Notably, however, under conditions in which hepatic cholesterol content was not enriched, activation of LXRs with the selective synthetic agonist GW3965 also acutely suppressed the expression of sterol synthesis genes in mouse liver (Fig. 1a and Extended Data Fig. 1a). The effect could not be explained by changes in intracellular cholesterol levels, as LXR activation has been shown to lower hepatic cholesterol content⁵, which would lead to upregulation of the SREBP-2 pathway.

To investigate the mechanism by which LXRs suppress cholesterol biosynthesis, we performed genome-wide transcriptional profiling on primary mouse hepatocytes treated with vehicle or GW3965 (Extended Data Fig. 1b). The most robustly induced gene in our RNA-sequencing (RNA-seq) analysis was a predicted non-coding RNA annotated as 4930412L05Rik (Extended Data Fig. 1c). Parallel profiling of non-coding and protein-coding transcripts using microarrays also identified 4930412L05Rik as the highest induced transcript (Extended Data Fig. 1d). We named this transcript *LeXis* (liver-expressed LXR-induced sequence). Notably, the *LeXis* gene locus lies in close proximity to the canonical LXR target gene *Abca1* in mouse. Analysis of chromatin structure from The ENCODE Project^{6,7} indicated that *LeXis* and *Abca1* were distinct genes with separate promoters

(Fig. 1b). We defined the transcripts produced from the *LeXis* gene using rapid amplification of complementary DNA ends (RACE) (Extended Data Fig. 2). *LeXis* and *Abca1* were induced by LXR and retinoid X receptor (RXR) agonists (LG268 and GW3965, respectively) in primary hepatocytes in an LXR-dependent manner (Fig. 1c and Extended Data Fig. 3a). *LeXis* was induced in *LXRα*^{-/-} and *LXRβ*^{-/-} (also known as *Nr1h3*^{-/-} and *Nr1h2*^{-/-}, respectively) hepatocytes, indicating that both LXR isotypes are capable of regulating *LeXis* (Extended Data Fig. 3b). Induction of *LeXis* was not sensitive to the protein synthesis inhibitor cycloheximide, and was not dependent on SREBPs, since 25-hydroxycholesterol (which blocks SREBP processing) also induced *LeXis* (Extended Data Fig. 3c, d).

Administration of GW3965 to mice induced the expression of *LeXis* in several metabolically active tissues (Fig. 1d and Extended Data Fig. 3e). We also observed a prominent, LXR-dependent induction of *LeXis* expression in response to Western diet feeding, consistent with a potential role for *LeXis* in the response to cholesterol excess (Fig. 1e). Despite being physically adjacent, the *LeXis* and *Abca1* loci are regulated independently. *LeXis* was neither expressed at baseline nor induced by LXR in mouse peritoneal macrophages, a cell type in which *Abca1* expression is prominent (Fig. 1f). A luciferase reporter containing the *LeXis* promoter was induced by LXR and RXR in co-transfection assays (Extended Data Fig. 3f), and we identified an LXR-response element within the *LeXis* promoter region that was bound by LXRα in chromatin immunoprecipitation and quantitative PCR (ChIP-qPCR) assays (Extended Data Fig. 3g). The coding potential calculator and coding-non-coding index algorithms predict low coding potential of *LeXis* (Extended Data Fig. 3h, i). In addition, we found no evidence of production of a protein product from *LeXis* using *in vitro* transcription-translation assays (Extended Data Fig. 3j).

To explore the function of *LeXis* *in vivo*, we transduced mice with adenoviral vectors encoding green fluorescent protein (GFP) control or *LeXis* (Fig. 2a and Extended Data Fig. 4a). Remarkably, *LeXis* expression decreased serum cholesterol, but not triglycerides, in chow-fed C57BL/6 mice (Fig. 2a, b). No differences in liver function tests were observed between the two groups, and there was no evidence of ER stress or inflammation (Fig. 2b and Extended Data Fig. 4b, c). Fractionation of lipoproteins revealed reduced cholesterol in both the low-density lipoprotein (LDL) and high-density lipoprotein (HDL) fractions in *LeXis*-expressing mice (Fig. 2c). The effects of *LeXis* were distinct from the consequences of hepatic expression of other LXR target genes, such as *Abca1* and *Idol* (also known as *Mylip*), which raise serum cholesterol^{8,9}.

Unbiased pathway analysis of global gene expression revealed that the cholesterol biosynthetic pathway was strongly downregulated in *LeXis*-transduced livers (Extended Data Fig. 4d). These results were validated by qPCR (Fig. 2d). These results suggested that the cholesterol lowering effects of *LeXis* were due, at least in part, to suppression

¹Department of Pathology and Laboratory Medicine, Howard Hughes Medical Institute, University of California, Los Angeles, California 90095, USA. ²Department of Medicine, Division of Cardiology, University of California, Los Angeles, California 90095, USA. ³Department of Human Genetics, University of California, Los Angeles, California 90095, USA. ⁴Ionis Pharmaceuticals, Carlsbad, California 92008, USA. ⁵Pasarrow Mass Spectrometry Laboratory, NPI-Semel Institute, University of California, Los Angeles, California 90095, USA.

*These authors contribute equally to this work.

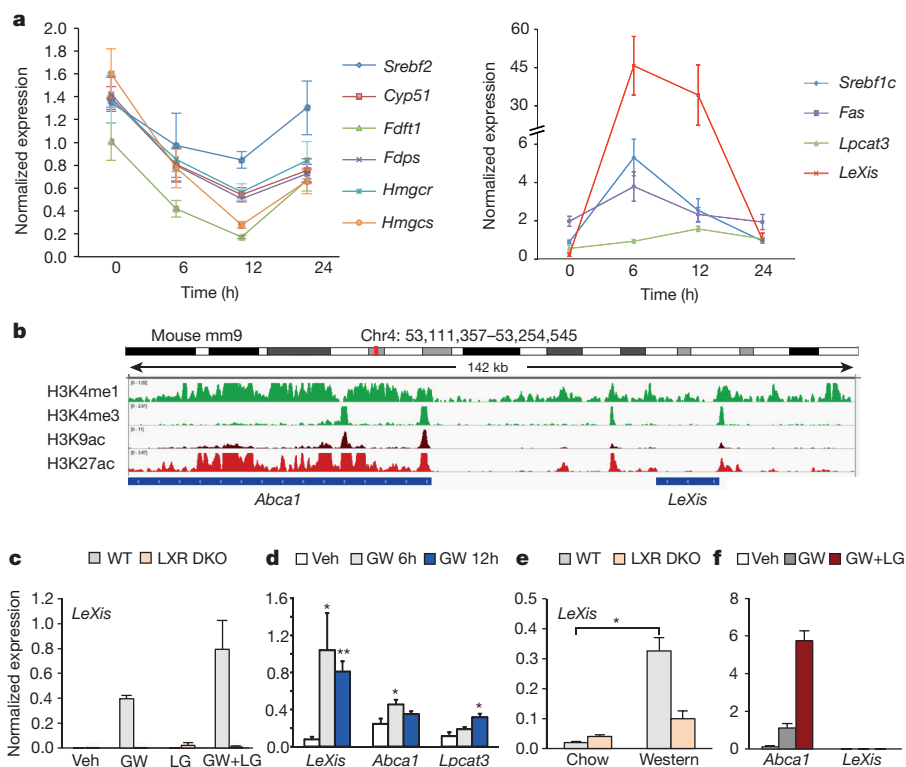


Figure 1 | LXR activation inhibits cholesterol biosynthesis and induces *LeXis* expression. **a**, qPCR analysis of gene expression in livers from C57BL/6 mice treated by oral gavage with 40 mg kg⁻¹ GW3965 for the indicated time ($n = 6$ mice per group). All curves are statistically different from baseline expression ($P < 0.05$, one way analysis of variance (ANOVA)). **b**, Schematic representation of the *LeXis* gene locus on an Integrative Genome Viewer (IGV) showing histone marks from LICR ENCODE data. **c**, qPCR analysis of gene expression in primary mouse hepatocytes treated with GW3965 (GW; 1 μM) and/or the RXR ligand LG268 (LG; 50 nM). DKO, double knockout (*LXRα*^{-/-} and *LXRβ*^{-/-}).

Results are representative of four independent experiments. **d**, qPCR analysis of gene expression in livers from male C57BL/6 mice gavaged with 40 mg kg⁻¹ GW3965 before collection at the indicated time ($n = 6$ per group). **e**, Gene expression in livers obtained from mice maintained on chow ($n = 2$ per group) or a Western diet ($n = 5$ per group). **f**, Gene expression in primary mouse peritoneal macrophages treated with 1 μM GW3965 and/or 50 nM LG268 for 16 h. Results are representative of four independent experiments. Values are mean \pm s.d. (**c**, **f**), or mean \pm s.e.m. (**a**, **d**, **e**). * $P < 0.05$; ** $P < 0.01$ (analysis of variance (ANOVA) with multi-group comparison in **a**, **d** and **e**).

of cholesterol biosynthesis. Consistent with this interpretation, we observed a strong trend towards lower cholesterol content in the livers of mice overexpressing *LeXis* (Extended Data Fig. 4e). For reasons that are not yet clear, treatment of isolated primary hepatocytes did not reflect the effects of either LXR agonist treatment or *LeXis* expression on genes linked to sterol synthesis (Extended Data Fig. 4f, g).

A reduction in plasma cholesterol suggests an increase in lipoprotein clearance or a decrease in sterol production¹⁰. To assess the contribution of the low-density lipoprotein receptor (LDLR) to the actions of *LeXis*, we transduced *Ldlr*^{-/-} mice with control or *LeXis*-expressing adenovirus. We observed decreases in plasma cholesterol levels and hepatic cholesterol content in response to *LeXis* in *Ldlr*^{-/-} mice, suggesting that the LDLR is not required for *LeXis* effects (Fig. 2e and Extended Data Fig. 4h, i). To assess the contribution of SREBP-2 signalling to LXR-mediated inhibition of cholesterologenesis, we administered GW3965 to control or liver-specific SCAP (L-SCAP) knockout mice¹¹. Consistent with previous studies¹², GW3965 treatment did not alter serum cholesterol levels in control mice (Fig. 2f). Notably, however, GW3965 increased serum cholesterol levels in L-SCAP knockout mice, suggesting the loss of a suppressive effect (Fig. 2f, g). LXR target genes, including *LeXis* itself, were induced by GW3965 in both groups; however, the suppression of steroidogenic genes was abrogated in L-SCAP knockout mice (Extended Data Fig. 4j). Furthermore, expression of *LeXis* also failed to lower serum cholesterol or suppress cholesterologenic gene expression in L-SCAP knockout mice (Fig. 2h and Extended Data Fig. 4k).

To address the role of *LeXis* in the setting of dietary cholesterol challenge, we used adenoviral vectors to express short hairpin RNA

(shRNA) constructs targeting *LeXis* in mouse liver^{13,14}. Knockdown of *LeXis* with either of two different shRNA constructs increased serum HDL cholesterol levels in mice fed a Western diet (Extended Data Fig. 5a–d). There was also an increase in liver cholesterol content in sh*LeXis*-transduced mice (Extended Data Fig. 5e). Gene expression analysis revealed increased expression of cholesterol biosynthetic genes in response to *LeXis* knockdown (Extended Data Fig. 5f). Similar effects of *LeXis* knockdown were observed in mice treated with GW3965 (Extended Data Fig. 5g, h). There was no consistent evidence of ER stress or inflammation in these experiments (Extended Data Fig. 5i–k).

As a complementary acute loss-of-function approach, we used antisense oligonucleotides (ASOs) to target *LeXis* expression^{15,16}. Three different ASOs that potentially blocked hepatic *LeXis*, but not saline or non-targeting ASO controls, increased serum cholesterol levels in the setting of LXR activation, with no evidence of hepatotoxicity (Fig. 3a, b and Extended Data Fig. 5l, m). Furthermore, *LeXis* ASO administration increased cholesterologenic gene expression (Fig. 3c).

We generated *LeXis*-deficient mice to determine the consequences of chronic loss of *LeXis* function (Extended Data Fig. 6a–c). Although serum cholesterol levels in *LeXis*-deficient mice in the setting of LXR activation were not different from controls (Fig. 3d), the expression of sterol synthesis genes in the liver was increased (Fig. 3e). Furthermore, *LeXis*-null mice had increased hepatic cholesterol content when challenged with a Western diet (Fig. 3f). Gross and histological examination of livers from *LeXis*-deficient null mice showed changes consistent with lipid accumulation (Fig. 3g, h). In contrast to the acute LXR agonist studies above, gene expression analysis of *LeXis*^{-/-} mice maintained

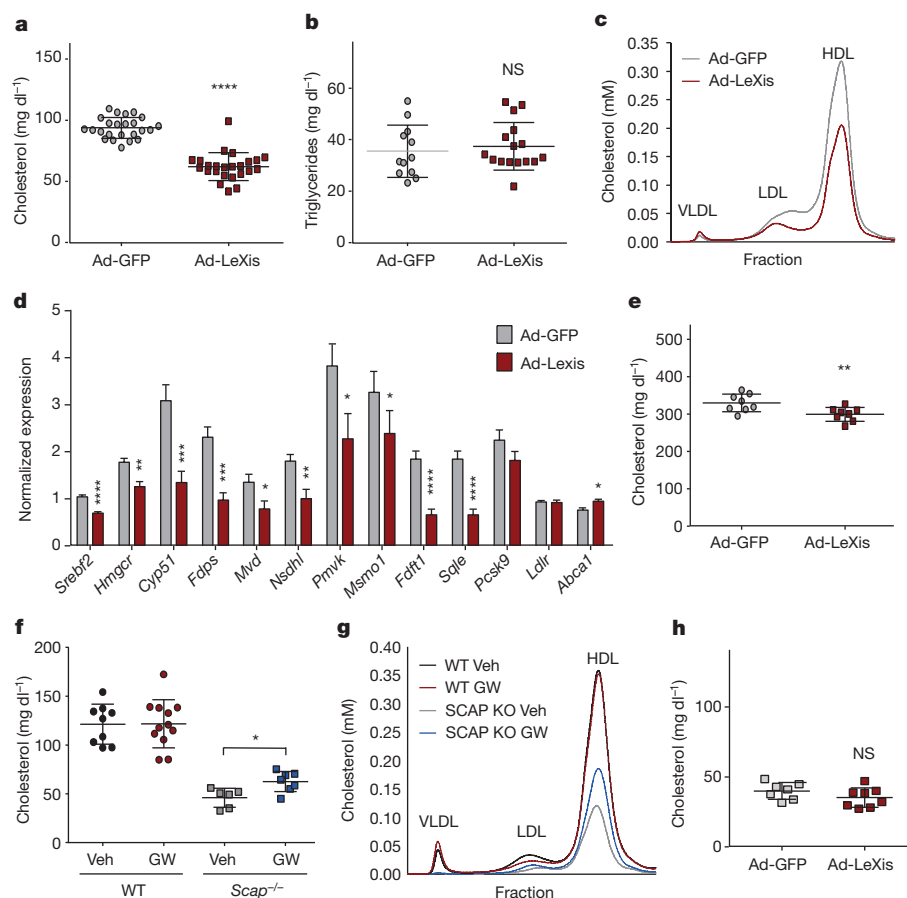


Figure 2 | *LeXis* expression reduces serum cholesterol and sterol synthesis through a pathway requiring intact SREBP signalling. **a**, Total serum cholesterol levels in 10-week-old chow-fed male C57BL/6 mice transduced with adenoviral vectors encoding GFP control (Ad-GFP) or *LeXis* (Ad-*LeXis*) for 6 days ($n = 24$ per group). **b**, Total serum triglycerides levels in the mice shown in **a** ($n = 12$ –16 per group). **c**, Cholesterol levels in pooled fractionated serum from mice treated with Ad-GFP or Ad-*LeXis*. VLDL, very-low-density lipoprotein. **d**, Analysis of gene expression in livers obtained after 6 days of transduction with Ad-GFP or Ad-*LeXis* ($n = 8$ per group). **e**, Total serum cholesterol levels in chow-fed male *Ldlr*^{-/-} mice (10 weeks old) transduced with Ad-GFP or Ad-*LeXis* for 6 days ($n = 8$ per group). **f**, Serum cholesterol levels in chow-fed wild-type (WT) or liver-specific SCAP knockout (*Scap*^{-/-}) mice gavaged with 40 mg kg⁻¹ GW3965 for 2 days. **g**, Cholesterol levels in pooled plasma fractions from mice shown in **f**. **h**, Total serum cholesterol levels in chow-fed *Scap*^{-/-} mice transduced with Ad-GFP or Ad-*LeXis* for 6 days. All values are mean \pm s.e.m. NS, not significant; * $P < 0.05$; ** $P < 0.01$; *** $P < 0.001$; **** $P < 0.0001$ (unpaired two-tailed *t*-test).

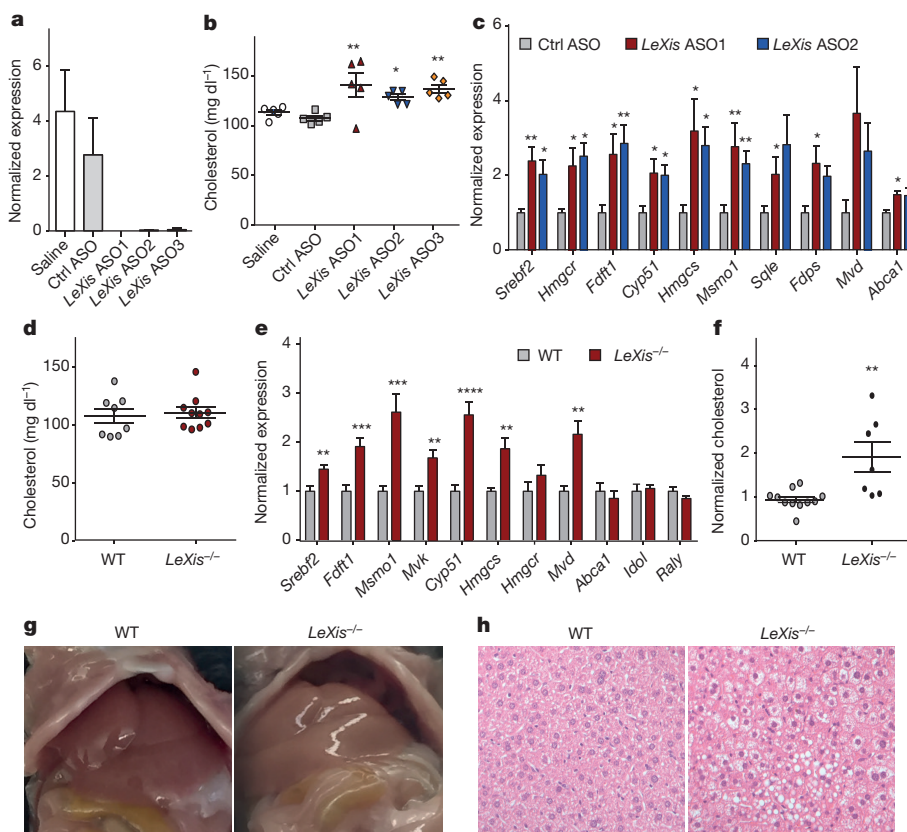


Figure 3 | Acute and chronic inactivation of *LeXis* alters hepatic lipid metabolism. **a**, *LeXis* gene expression (normalized to 36B4, also known as *Rplp0*) in livers from C57BL/6 mice on a chow diet administered 25 mg kg⁻¹ ASOs intraperitoneally on days 1, 4 and 7, and gavaged with 40 mg kg⁻¹ GW3965 on days 4, 7 and 8 ($n = 5$ per group). **b**, Total serum cholesterol from mice in **a**. **c**, Gene expression from C57BL/6 mice on a chow diet administered 25 mg kg⁻¹ ASOs intraperitoneally on days 1, 3 and 5, and gavaged with 40 mg kg⁻¹ GW3965 on days 5 and 6 ($n = 8$ per group). **d**, Total serum cholesterol levels in chow-fed wild-type or *LeXis*^{-/-} mice gavaged with 40 mg kg⁻¹ GW3965 for 2 days ($n = 8$ –10 per group). **e**, Gene expression from C57BL/6 wild-type or *LeXis*^{-/-} mice on a chow diet gavaged with 40 mg kg⁻¹ GW3965 for 2 days ($n = 8$ –10 per group). **f**, Hepatic cholesterol content was normalized to liver mass from C57BL/6 wild-type or *LeXis*^{-/-} mice fed a Western diet for 3 weeks ($n = 7$ –11 per group). **g**, Representative (of three images per group) gross appearance of livers from wild-type and *LeXis*^{-/-} mice after 3 weeks on a Western diet. **h**, Histological sections of liver from wild-type and *LeXis*^{-/-} mice after 3 weeks on a Western diet (haematoxylin and eosin stain representative of three images per group). Original magnifications, $\times 40$ (**g**) and $\times 1$ (**h**). All values (**a**–**f**) are mean \pm s.e.m. * $P < 0.05$; ** $P < 0.01$; *** $P < 0.001$; **** $P < 0.0001$ (ANOVA (**b**, **c**) and unpaired two-tailed *t*-test (**e**, **f**)).

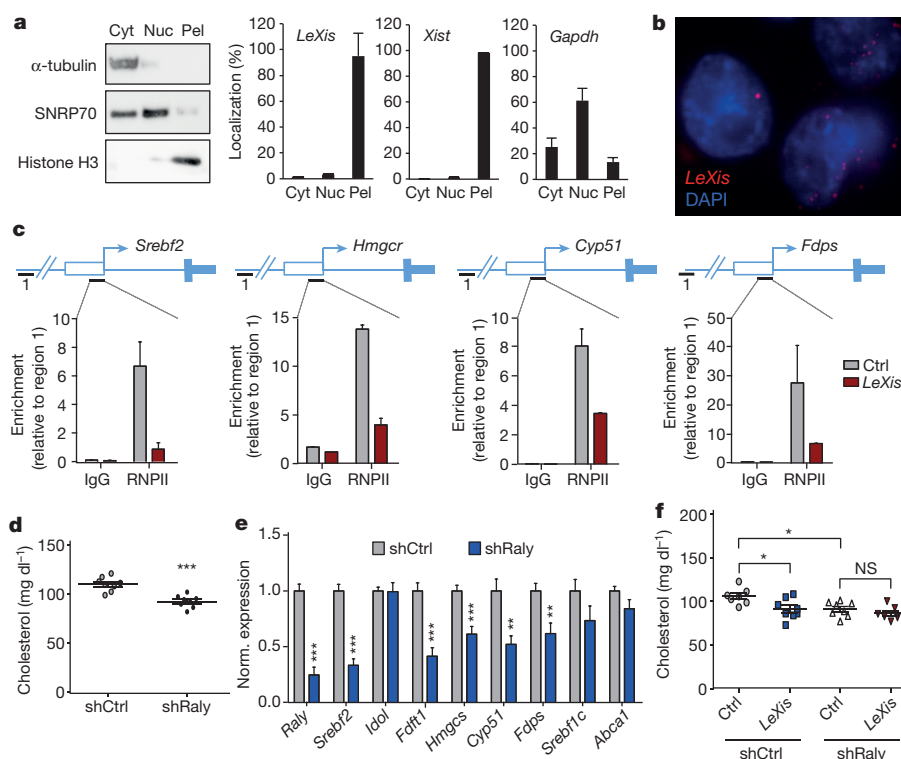


Figure 4 | *LeXis* interacts with RALY to regulate metabolic gene expression. **a**, Hepa1-6 cells were transfected with *LeXis*, and 24 h later cellular content was separated into cytoplasmic soluble (cyt), nuclear soluble (nuc) and insoluble pellet (pel) fractions. Transcripts in each fraction were analysed by qPCR, and fraction purity was validated by western blotting with the indicated compartment markers ($n = 3$ per group). **b**, Representative (of three) micrograph showing *LeXis* subcellular localization in primary mouse hepatocytes by single molecule fluorescence *in situ* hybridization using anti-sense probes to *LeXis* (red). Nuclei were counterstained with DAPI (blue). Original magnification, $\times 63$. **c**, Recruitment of RNA polymerase II (RNPII) to promoter regions as determined by ChIP-qPCR analysis in livers transduced with control

(Ad-GFP) or *LeXis*-expressing (Ad-*LeXis*) adenoviruses. Data are expressed as percentage input retrieved normalized to an upstream site (region 1) ($n = 3$ per group). **d**, Total serum cholesterol levels in 14-week-old chow-fed male C57BL/6 mice transduced with control (shCtrl) or adenoviral vectors expressing *Raly* shRNA (shRaly) ($n = 8$ per group). **e**, Gene expression in livers of the mice shown in **f**. **f**, Total serum cholesterol in chow-fed male C57BL/6 mice transduced with control (Ad-GFP) or Ad-*LeXis* (1.0×10^9 plaque-forming units, p.f.u.) and shCtrl or shRaly (2.0×10^9 p.f.u.) ($n = 7-8$ per group). Values are mean \pm s.d. (**a**, **c**) or mean \pm s.e.m. (**d**). * $P < 0.05$; ** $P < 0.01$; *** $P < 0.001$ (unpaired two-tailed *t*-test (**d**, **e**) and ANOVA with multi-group comparison (**f**)).

on Western diet showed a trend towards decreased sterol synthetic gene expression, probably reflecting the marked increased hepatic cholesterol content in this setting (Extended Data Fig. 6d).

To begin to understand how *LeXis* was influencing hepatic metabolism, we analysed its subcellular localization. *LeXis* was almost exclusively located in the insoluble nuclear pellet in fractionation studies, along with the known nuclear long non-coding RNAs (lncRNAs) *XIST* and histone H3 (Fig. 4a). Single molecule RNA fluorescence *in situ* hybridization with *LeXis*-specific probes further confirmed its nuclear localization (Fig. 4b).

Owing to the presence of *LeXis* in the nucleus, we tested its ability to affect RNA polymerase II-dependent transcription. Expression of *LeXis* in mouse liver reduced RNA polymerase II engagement at the promoters of *Srebf2* and its target genes (Fig. 4c). Previous work has shown that nuclear lncRNAs can affect transcription by modifying the recruitment of proteins to chromatin¹⁷. We used an unbiased lncRNA-chromatin affinity capture technique to pull-down *LeXis* from mouse liver and identify interacting proteins¹⁸ (Extended Data Fig. 7a). Analysis of the *LeXis* interactome by mass spectrometry identified the heterogeneous ribonucleoprotein RALY¹⁹ as a binding partner. Similar to *LeXis*, RALY was located in the nuclear pellet (chromatin) fraction of hepatocytes (Extended Data Fig. 7b). Moreover, an antibody to RALY retrieved *LeXis* in co-immunoprecipitation studies (Extended Data Fig. 7c, d).

RALY contains both an RNA-binding domain and a leucine-zipper coiled domain, suggesting it may act as a regulatory factor²⁰.

Notably, previous unbiased analysis of gene coexpression networks has identified *Srebf2* as one of the top genes positively coregulated with *Raly*²¹. Other studies have shown direct binding of SREBP-2 at the *Raly* promoter²². Unbiased protein homology analysis revealed extensive structural conservation between RALY and RNA binding motif protein 14 (RBM14, also known as CoAA)²³ (Extended Data Fig. 7e), a known steroid receptor coactivator²⁴. This led us to hypothesize that RALY may act as transcriptional cofactor for genes involved in cholesterol biosynthesis. In line with this idea, adenovirus-mediated knockdown of RALY in mouse liver reduced serum cholesterol, mimicking the effect of *LeXis* expression (Fig. 4d and Extended Data Fig. 7f). This effect was correlated with reduced expression of *Srebf2* and its target genes (Fig. 4e and Extended Data Fig. 7g). Unbiased gene expression profiling of liver revealed that RALY knockdown preferentially affected cholesterol biosynthetic pathways (Extended Data Fig. 8a, b). The effects of RALY were independent of LDLR expression, since they were preserved in *Ldlr*-null mice (Extended Data Fig. 9a, b). The actions of *LeXis* *in vivo* were dependent on RALY, since the ability of *LeXis* to alter serum cholesterol levels and hepatic gene expression was impaired in the setting of RALY knockdown (Fig. 4f and Extended Data Fig. 9c). Finally, ChIP-qPCR analysis of mouse liver revealed that RALY associated with cholesterol biosynthetic gene promoters, and that RALY occupancy was reduced in the setting of *LeXis* expression (Extended Data Fig. 9d).

This work identifies the non-coding RNA *LeXis* as an additional mediator of the complex effects of LXR signalling on hepatic lipid

metabolism. Our data suggest that *LeXis* contributes to the ability of LXRs to inhibit cholesterol synthesis. It is important to acknowledge, however, that the involvement of additional pathways in this crosstalk is not excluded by the present work. The demonstration that *LeXis* expression is responsive to dietary cues and can modulate physiological pathways with links to common diseases expands our understanding of the regulatory potential of non-coding RNA. Notably, the consequences of acute and chronic loss of *LeXis* expression are only partially overlapping, perhaps reflecting compensation in the setting of developmental deletion²⁵.

Although the rapid sequence evolution of lncRNAs presents a challenge to identifying functional counterparts between species²⁶, batch coordinate conversion between mouse and human assemblies revealed moderate conservation of the *LeXis* genomic sequence in a region adjacent to the human *ABCA1* gene. An annotated putative lncRNA (TCONS_00016452) in this region was robustly induced by LXR activation in human hepatocyte cell lines (Extended Data Fig. 10). In the future it will be of interest to assess whether this sequence or an as yet to be identified lncRNA is a functional orthologue of *LeXis*.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 3 August 2015; accepted 18 March 2016.

Published online 11 May 2016.

1. Tontonoz, P. Transcriptional and posttranscriptional control of cholesterol homeostasis by liver X receptors. *Cold Spring Harb. Symp. Quant. Biol.* **76**, 129–137 (2011).
2. Repa, J. J. *et al.* Regulation of mouse sterol regulatory element-binding protein-1c gene (SREBP-1c) by oxysterol receptors, LXR α and LXR β . *Genes Dev.* **14**, 2819–2830 (2000).
3. Zelcer, N., Hong, C., Boyadjan, R. & Tontonoz, P. LXR regulates cholesterol uptake through Idol-dependent ubiquitination of the LDL receptor. *Science* **325**, 100–104 (2009).
4. Brown, M. S. & Goldstein, J. L. The SREBP pathway: regulation of cholesterol metabolism by proteolysis of a membrane-bound transcription factor. *Cell* **89**, 331–340 (1997).
5. Zhang, Y. *et al.* Liver LXR α expression is crucial for whole body cholesterol homeostasis and reverse cholesterol transport in mice. *J. Clin. Invest.* **122**, 1688–1699 (2012).
6. Creighton, M. P. *et al.* Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proc. Natl Acad. Sci. USA* **107**, 21931–21936 (2010).
7. The ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
8. Zelcer, N. & Tontonoz, P. Liver X receptors as integrators of metabolic and inflammatory signaling. *J. Clin. Invest.* **116**, 607–614 (2006).
9. Vaisman, B. L. *et al.* ABCA1 overexpression leads to hyperalphalipoproteinemia and increased biliary cholesterol excretion in transgenic mice. *J. Clin. Invest.* **108**, 303–309 (2001).
10. Horton, J. D., Goldstein, J. L. & Brown, M. S. SREBPs: activators of the complete program of cholesterol and fatty acid synthesis in the liver. *J. Clin. Invest.* **109**, 1125–1131 (2002).
11. Matsuda, M. *et al.* SREBP cleavage-activating protein (SCAP) is required for increased lipid synthesis in liver induced by cholesterol deprivation and insulin elevation. *Genes Dev.* **15**, 1206–1216 (2001).
12. Hong, C. *et al.* The LXR-Idol axis differentially regulates plasma LDL levels in primates and mice. *Cell Metab.* **20**, 910–918 (2014).
13. Carpenter, S. *et al.* A long noncoding RNA mediates both activation and repression of immune response genes. *Science* **341**, 789–792 (2013).
14. Yang, F., Zhang, H., Mei, Y. & Wu, M. Reciprocal regulation of HIF-1 α and lincRNA-p21 modulates the Warburg effect. *Mol. Cell* **53**, 88–100 (2014).
15. Raal, F. J. *et al.* Mipomersen, an apolipoprotein B synthesis inhibitor, for lowering of LDL cholesterol concentrations in patients with homozygous familial hypercholesterolaemia: a randomised, double-blind, placebo-controlled trial. *Lancet* **375**, 998–1006 (2010).
16. Gaudet, D. *et al.* Targeting APOC3 in the familial chylomicronemia syndrome. *N. Engl. J. Med.* **371**, 2200–2206 (2014).
17. Rinn, J. L. & Chang, H. Y. Genome regulation by long noncoding RNAs. *Annu. Rev. Biochem.* **81**, 145–166 (2012).
18. Chu, C., Quinn, J. & Chang, H. Y. Chromatin isolation by RNA purification (ChIRP). *J. Vis. Exp.* **61**, 3912 (2012).
19. Michaud, E. J., Bultman, S. J., Stubbs, L. J. & Woychik, R. P. The embryonic lethality of homozygous lethal yellow mice (*A^y/A^y*) is associated with the disruption of a novel RNA-binding protein. *Genes Dev.* **7**, 1203–1213 (1993).
20. Jiang, W., Guo, X. & Bhavanandan, V. P. Four distinct regions in the auxiliary domain of heterogeneous nuclear ribonucleoprotein C-related proteins. *Biochim. Biophys. Acta* **1399**, 229–233 (1998).
21. Okamura, Y. *et al.* COXPRESdb in 2015: coexpression database for animal species by DNA-microarray and RNAseq-based expression data with multiple quality assessment systems. *Nucleic Acids Res.* **43**, D82–D86 (2015).
22. Seo, Y. K. *et al.* Genome-wide localization of SREBP-2 in hepatic chromatin predicts a role in autophagy. *Cell Metab.* **13**, 367–375 (2011).
23. Kelley, L. A., Mezulis, S., Yates, C. M., Wass, M. N. & Sternberg, M. J. The Phyre2 web portal for protein modeling, prediction and analysis. *Nature Protocols* **10**, 845–858 (2015).
24. Auboeuf, D. *et al.* CoAA, a nuclear receptor coactivator protein at the interface of transcriptional coactivation and RNA splicing. *Mol. Cell. Biol.* **24**, 442–453 (2004).
25. Rossi, A. *et al.* Genetic compensation induced by deleterious mutations but not gene knockdowns. *Nature* **524**, 230–233 (2015).
26. Ulitsky, I., Shkumatava, A., Jan, C. H., Sive, H. & Bartel, D. P. Conserved function of lincRNAs in vertebrate embryonic development despite rapid sequence evolution. *Cell* **147**, 1537–1550 (2011).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank members of the Tontonoz, Nagy, Smale and Black laboratories and the UCLA Atherosclerosis Research Unit for technical assistance and useful discussions. This work was supported by NIH grants HL030568, HL066088, DK063491, HL128822, DK102559 and HL69766; American Heart Association grant 13POST17080115; American College of Cardiology Presidential CDA; and the UCLA Cardiovascular Discovery Fund (Lauren B. Leichtman and Arthur E. Levine Investigator Award).

Author Contributions T.S. and P.T. conceived and designed the study, guided the interpretation of the results and the preparation of the manuscript. P.T. supervised the study and provided critical suggestions. T.S. and X.W. performed most mouse experiments and data analysis. M.C.J., T.G., L.Z., J.S., C.H., T.d.A.V. participated in mouse experiments and data analysis. T.S. performed RNA-seq experiments and validated *LeXis* as an LXR target. A.E. and D.C. processed and analysed next-generation sequencing data. M.C.J. performed and analysed the RACE experiments. J.W. performed the mass spectrometry analysis. M.K. and R.L. provided and independently validated ASOs targeting *LeXis*. T.S. and P.T. drafted the manuscript. T.S., M.C.J. and P.T. edited the manuscript with input from all authors. All authors discussed the results and approved the final version of the manuscript.

Author Information Sequencing and microarray data have been deposited in the Gene Expression Omnibus (GEO) under accessions GSE77793, GSE77786, GSE77802 and GSE77805. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to P.T. (ptontonoz@mednet.ucla.edu).

METHODS

Reagents, plasmids and gene expression. GW3965 was synthesized as previously described²⁷. LG268 was from Ligand Pharmaceuticals. Oxysterols were purchased from Sigma and used as described²⁸. Simvastatin sodium salt was from Calbiochem. Ligands were dissolved in dimethyl sulfoxide before use in cell culture. *LeXis* was amplified from GW3695-treated primary mouse hepatocytes using KOD polymerase (Millipore) and primers designed to provide flanking attB sequences and a SacI site at the immediate 3' end. The fragments were then cloned into pDONR221 using the Gateway system and the minimal SV40 polyadenylation sequence was inserted at the SacI site. For transient transfections and viral vector production the entry clone was transferred into the pAd/CMV/V5-DEST Gateway vector by LR recombination. We estimate transcription from this vector to append 109 nucleotides at the 5' end, and 29 nucleotides at the 3' end of the cloned *LeXis* sequence. To obtain the sh*LeXis* adenovirus, we used BLOCK-iT kit as described (Invitrogen)³. In brief, Invitrogen based software was used for original nucleotide generation targeting the LEXIS fragment and cloned into pENTR/U6. The resulting pENTR/U6-LEXIS shRNA plasmids were tested for their ability to inhibit overexpressed LEXIS in transient transfection experiments in HEK293T cells and then transferred by Gateway recombination into the pAd/BLOCK-iT-DEST destination vector for viral particle generation. Viruses were amplified, purified and titred by Viraquest. For gene expression analysis, RNA was isolated using TRIzol reagent (Invitrogen) and analysed by qPCR using an Applied Biosystems 7900HT sequence detector or Applied Biosystems Quant Studio 6 Flex. Results are normalized to 36B4 or cyclophilin (also known as *Ppia*). Immunohistochemical staining of paraffin-embedded livers were done by the UCLA Translational Pathology Core Laboratory.

Animals and diets. All animals (C57BL/6, greater than 10 generations backcrossed) were housed in a temperature-controlled room under a 12-h light/12-h dark cycle and pathogen-free conditions. For adenovirus experiments, age-matched mice were purchased from Jackson Laboratories. Littermates were manually randomized to different treatment groups. Investigators were blinded to group allocation for some but not all studies. *LXRα*^{-/-}, *LXRβ*^{-/-} and *LXRαβ*^{-/-} mice were originally provided by D. Mangelsdorf. Floxed *Scap*^{-/-} mice were previously described²⁹. *LeXis* global knockout mice were generated at UC Davis KOMP using strategy outlined in Extended Data Fig. 6. Mice were fed a chow diet except as indicated, where mice were placed on a Western diet (21% fat, 0.21% cholesterol; D12079B; Research Diets Inc.) or were gavaged with either vehicle or 40 mg kg⁻¹ GW3965. Livers were obtained 4 h after the last gavage. We measured cholesterol and triglycerides as previously described³⁰. For adenoviral infections, age-matched (9–11 weeks old) male mice were injected with 2.0×10^9 p.f.u. by tail-vein injection unless otherwise specified. Mice were euthanized 6 days later after a 6-h fast. At the time of euthanization, liver tissue and blood was collected by cardiac puncture and immediately frozen in liquid nitrogen and stored at -80°C. Liver tissue was processed for isolation of RNA and protein as above. Generation 2.5 constrained ethyl ASOs, synthesized as described previously³¹, were administered by three 25 mg kg⁻¹ intraperitoneal doses together with 40 mg kg⁻¹ GW3965. Animals were euthanized on day 6 or 8 as indicated in figure legends. Most experiments were performed using male mice. All animal experiments were approved by the UCLA Institutional Animal Care and Research Advisory Committee.

Cell culture. Primary peritoneal macrophages were isolated 4 days after thioglycollate injection and prepared as described³². Mouse primary hepatocytes were isolated as previously described and cultured in William's E medium with 5% FBS²⁸. Peritoneal cells were incubated in 0.5% FBS in DMEM, with 5 μM simvastatin and 100 μM mevalonic acid. Five to eight hours later, cells were pretreated with dimethylsulfoxide (DMSO) or appropriate ligand overnight. *In vitro* translation assay was performed using TnT Coupled Transcription/Translation System (PROMEGA) according to the manufacturer's protocol. The cell lines HEK293T, HEK293A and Hepa1-6 were originally obtained from ATCC. All cells were tested for mycoplasma contamination.

RACE. The 5' and 3' ends of the *LeXis* transcript were defined using mouse liver RNA and the FirstChoice RLM-RACE kit (Ambion) according to manufacturer's protocol, with modifications. In brief, for the 5' RACE, degraded messenger RNA 5' ends were dephosphorylated with CIP, and then full-length mRNA was decapped with TAP. Following 5' RACE adaptor ligation, reverse transcription was performed using SuperScriptIII First-Strand Synthesis system (Invitrogen) and *LeXis*-specific primers. For the 3' RACE, RNA was reverse transcribed using SuperScriptIII First-Strand Synthesis system (Invitrogen) and the adaptor-linked oligo dTs. The resulting cDNA was amplified by nested PCR across a 55–65°C melting temperature gradient using KOD polymerase (Millipore), with the inner primers containing attB sequences. Aliquots of reactions were inspected on 1% agarose gels for product size and abundance. Products of select PCR reactions were purified using NucleoSpin Gel and PCR Cleanup kit (Clontech) and were inserted

into pDONR221 by Gateway cloning. Cloned fragments were sequenced and then aligned to the mouse genome with the BLAST analysis tool.

RNA fractionation. The Pad/CMV-*LeXis* vector was transfected into Hepa1-6 cells using BioT reagent (Bioland Scientific LLC) and 24 h later subcellular RNA fractions were obtained according to the protocol described previously³³. Lysate aliquots were inspected for fractional purity by western blotting with antibodies against α-tubulin, SNRP70 and histone H3 as cytoplasmic, nucleoplasmic and chromatin bound markers, respectively.

RNA-seq. RNA-seq libraries, starting with 500 ng total RNA, were constructed with the TruSeq RNA Sample Prep Kits from Illumina on RNA isolated from primary hepatocytes treated with or without GW3965. Samples were indexed with adapters and submitted for paired-end 2 × 100-bp sequencing in Illumina HiSeq2000. RNA-seq reads were aligned with TopHatv2.0.2 to the mouse genome, version mm9 (ref. 34). The TopHat alignment rate was 85%, resulting in an average of 65 million reads per sample. Transcripts were assessed and quantities were determined by Cufflinks v2.0.2, using a GTF file based on Ensembl mouse NCBI37. Comparison expression levels were made using fragments per kilobase of exon per million fragments mapped (FPKM) values using Cuffdiff from the Cufflinks package³⁵. Data analysis was performed by UCLA DNA Microarray Core.

Lipid analysis. Tissue lipid was obtained using a Folch extraction. In brief, chloroform extracts were dried under nitrogen and solubilized in water. Tissue and serum cholesterol and triglycerides were determined using a commercially available enzymatic kit (Wako). Hepatic cholesterol content was normalized to liver weight and protein concentration. Mice were fasted for at least 6 h before blood collection and euthanization. Plasma lipoprotein fractions were analysed by FPLC.

Microarray. For cDNA microarray analysis, primary hepatocytes cells were treated as indicated above with either DMSO or GW3965. These samples were from an independent cohort from those submitted for RNA-seq. For each condition, two independent samples were processed. Transcriptional profiling was performed at the University of California, Los Angeles, microarray core facility by using Agilent SurePrint G3 Gene Expression array. Data were analysed using GeneSpring software (Agilent Technologies) and David³⁶.

ChIP. ChIP studies were performed as described elsewhere³⁷. In brief, mouse livers were cross-linked using a final formaldehyde concentration of 1% at room temperature for 10 min. The reaction was quenched with the addition of glycine. For sonication, 0.3 ml (1/3) of nuclear lysate was sonicated for 25–30 cycles, 30 s on 30 s off at 4°C, with BioRuptor twin sonicator (Diagenode). Sonicated Chromatin was incubated overnight at 4°C with control IgG or 25 μg of anti-LXRα antibody (PPZ0412, ChIP Ggade, Abcam), anti-RALY antibody (EPRI0121, Abcam), or Pol II antibody (N-20, Santa Cruz Biotechnology). Protein A dynabeads (50 μl per immunoprecipitation sample) were added for 4 h. After incubation beads were washed with wash buffer A (50 mM HEPES, pH 7.9, 140 mM NaCl, 1 mM EDTA, 1% Triton X-100, 0.1% Na-deoxycholate, 0.1% SDS, 1× protease inhibitors freshly added), buffer B (50 mM HEPES, pH 7.9, 500 mM NaCl, 1 mM EDTA, 1% Triton X-100, 0.1% Na-deoxycholate, 0.1% SDS) and finally LiCL buffer (20 mM Tris, pH 8.0, 250 mM LiCl, 1 mM EDTA, 0.5% Na-deoxycholate, 0.5% NP-40). Reverse crosslinking was performed at 60°C overnight, mixed at 1,000 r.p.m., and DNA was extracted using a phenol–chloroform phase lock tube (5 PRIME) or Nucleospin PCR cleanup column (Macherey-Nagel). A standard curve for PCR was generated from serial dilutions of input samples and data expressed as percentage of input.

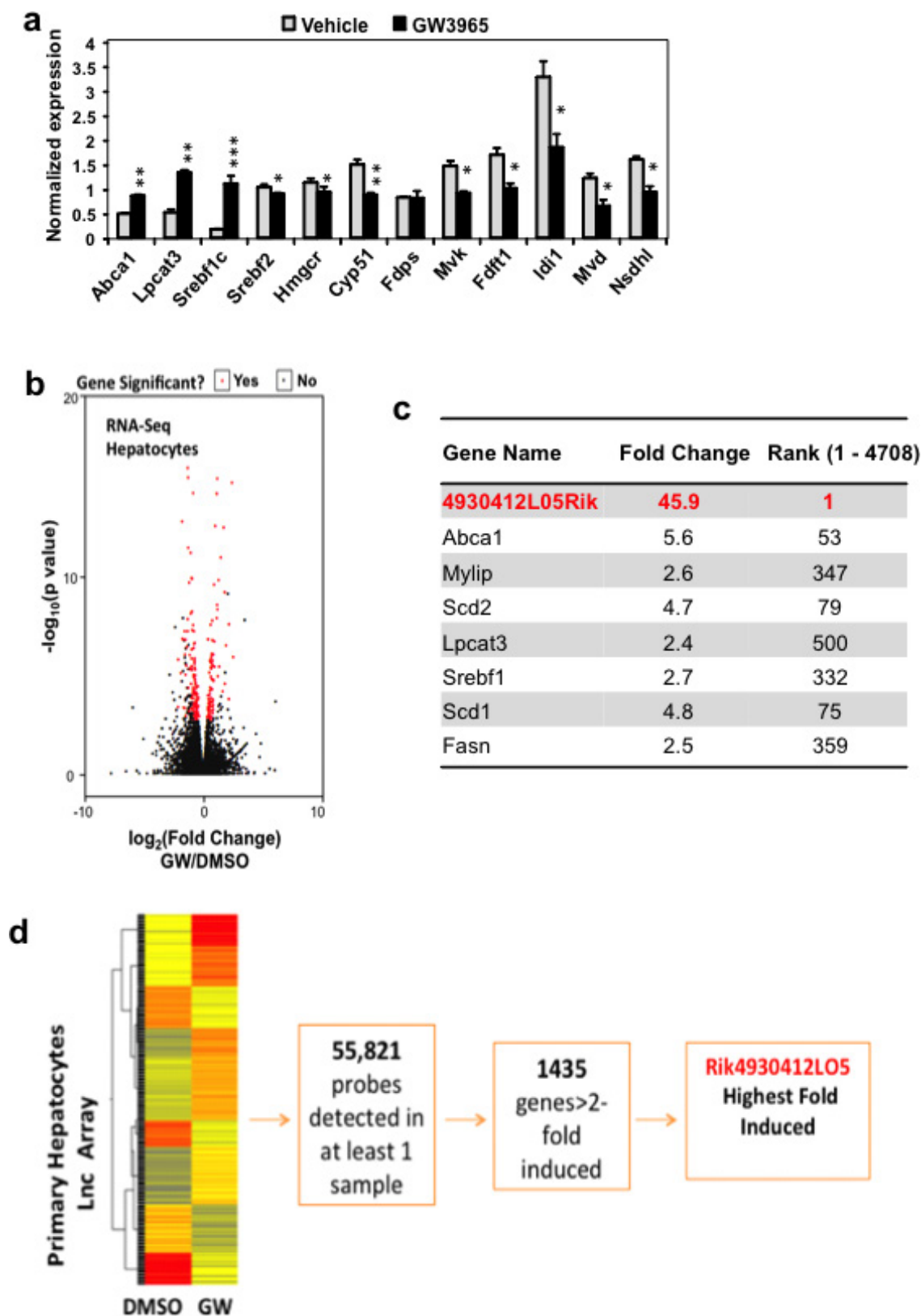
Chromatin isolation by RNA purification. Chromatin isolation by RNA purification (ChIRP) was performed as described previously¹⁸. In brief, mouse livers were cross-linked using glutaraldehyde. After glycine quenching, the nuclear lysate was sonicated for 25–30 cycles, 30 s on 30 s off at 4°C, with BioRuptor twin sonicator (Diagenode). *LeXis* and LacZ pull-down probes with BiotinTEG at 3' were designed by Biosearch Technologies (see Supplementary Information) and allowed to hybridize overnight with sonicated chromatin at 37°C (100 pmol probe per 1 ml chromatin). After hybridization, C1 Dynabeads (Life Technologies) were added and incubated for 30 min. For protein elution for mass spectrometry analysis, washed beads were resuspended in 3× original volume of DNase buffer (100 mM NaCl and 0.1% NP-40), and protein was eluted with a cocktail of 50 mM triethyl ammonium bicarbonate, 12 mM sodium lauryl sarcosine, and 0.5% sodium deoxycholate supplemented with 100 μg ml⁻¹ RNase A (Sigma-Aldrich) and 0.1 U μl⁻¹ RNase H (Epicentre), and 100 U ml⁻¹ DNase I (Invitrogen). For RNA isolation, beads were resuspended in proteinase K buffer (100 mM NaCl, 10 mM TrisCl, pH 7.0, 1 mM EDTA, 0.5% SDS, 5% by volume proteinase K (AM2546, Ambion) 20 mg ml⁻¹) and incubated at 50°C followed by Trizol isolation and DNase treatment.

Single molecule RNA FISH. Custom Stellaris FISH probes were designed against *LeXis*. Stellaris probe set labelled with CAL Fluor Red 610 and RNA FISH performed as described previously³⁸. In brief, hepatocytes were fixed with 3.7%

formaldehyde in PBS followed by 70% ethanol treatment to permeabilize cells. Cells were washed with 10% formamide in 2× SSC followed by treatment in humidified chamber with addition of probes (125 nM) in hybridization buffer (100 mg ml⁻¹ dextran sulfate and 10% formamide in 2× SSC). Cells were incubated in the dark at 37°C for 4 h. DAPI nuclear stain (5 ng ml⁻¹) was applied after washing with 10% formamide in 2× SSC. Images obtained using a Zeiss Z1 AxioObserver fluorescent microscope.

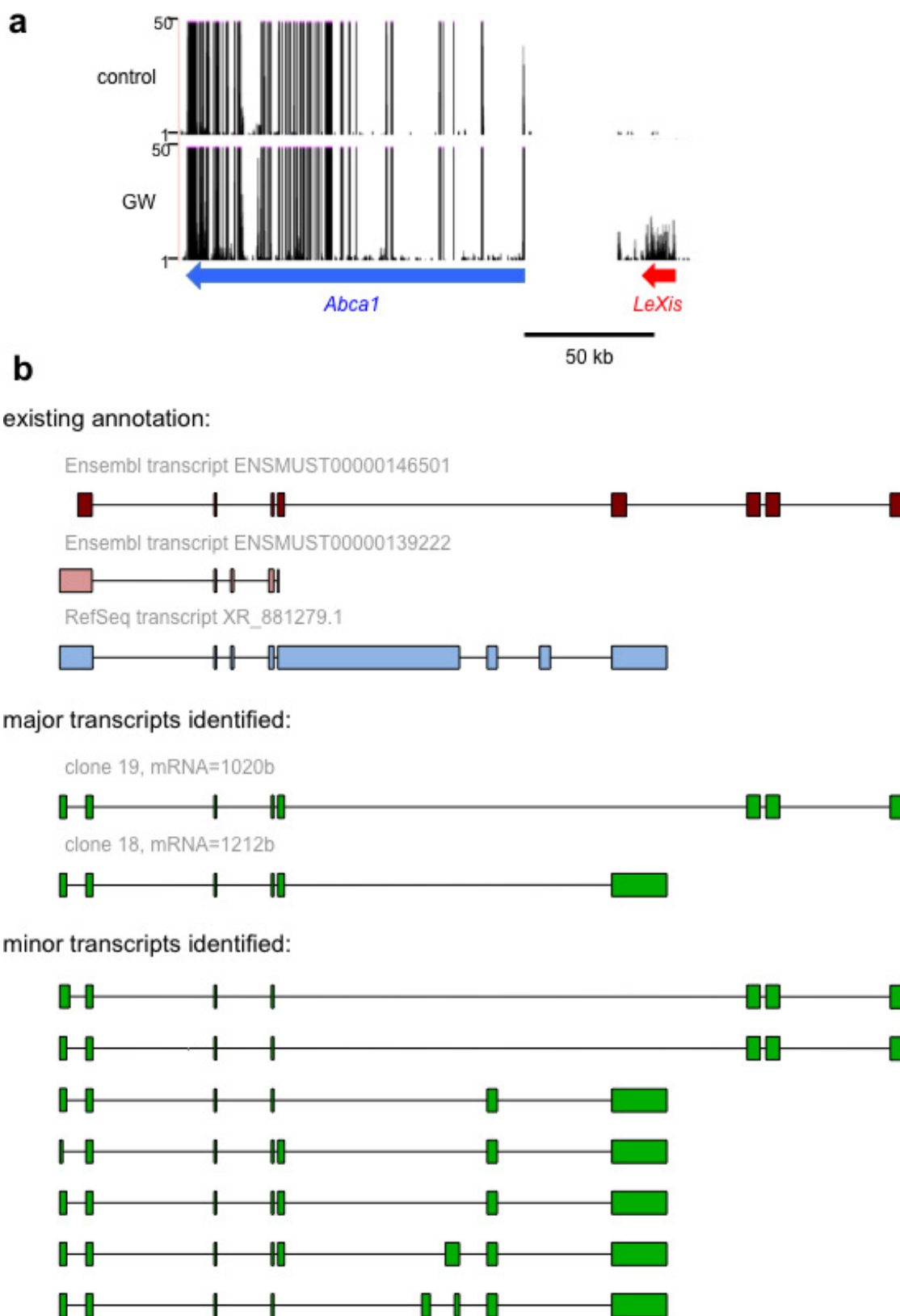
Statistical analysis. A non-paired Student's *t*-test or ANOVA was used to determine statistical significance, defined at *P* < 0.05. Unless otherwise noted, error bars represent s.d.. Experiments were independently performed at least twice. Group sizes were based on statistical analysis of variance and prior experience with similar *in vivo* studies.

27. Sallam, T. *et al.* The macrophage LBP gene is an LXR target that promotes macrophage survival and atherosclerosis. *J. Lipid Res.* **55**, 1120–1130 (2014).
28. Rong, X. *et al.* LXRs regulate ER stress and inflammation through dynamic modulation of membrane phospholipid composition. *Cell Metab.* **18**, 685–697 (2013).
29. Tarling, E. J., Ahn, H. & de Aguiar Vallim, T. Q. The nuclear receptor FXR uncouples the actions of miR-33 from SREBP-2. *Arterioscler. Thromb. Vasc. Biol.* **35**, 787–795 (2015).
30. Hong, C. *et al.* LXR α is uniquely required for maximal reverse cholesterol transport and atheroprotection in ApoE-deficient mice. *J. Lipid Res.* **53**, 1126–1133 (2012).
31. Seth, P. P. *et al.* Short antisense oligonucleotides with novel 2'–4' conformationally restricted nucleoside analogues show improved potency without increased toxicity in animals. *J. Med. Chem.* **52**, 10–13 (2009).
32. Bradley, M. N. *et al.* Ligand activation of LXR β reverses atherosclerosis and cellular cholesterol overload in mice lacking LXR α and apoE. *J. Clin. Invest.* **117**, 2337–2346 (2007).
33. Bhatt, D. M. *et al.* Transcript dynamics of proinflammatory genes revealed by sequence analysis of subcellular RNA fractions. *Cell* **150**, 279–290 (2012).
34. Trapnell, C., Pachter, L. & Salzberg, S. L. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25**, 1105–1111 (2009).
35. Trapnell, C. *et al.* Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnol.* **28**, 511–515 (2010).
36. Huang, D. W., Sherman, B. T. & Lempicki, R. A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature Protocols* **4**, 44–57 (2009).
37. Carey, M. F., Peterson, C. L. & Smale, S. T. Chromatin immunoprecipitation (ChIP). *Cold Spring Harb. Protoc.* **2009**, pdb.prot5279 (2009).
38. Raj, A. & Tyagi, S. Detection of individual endogenous RNA transcripts in situ using multiple singly labeled probes. *Methods Enzymol.* **472**, 365–386 (2010).



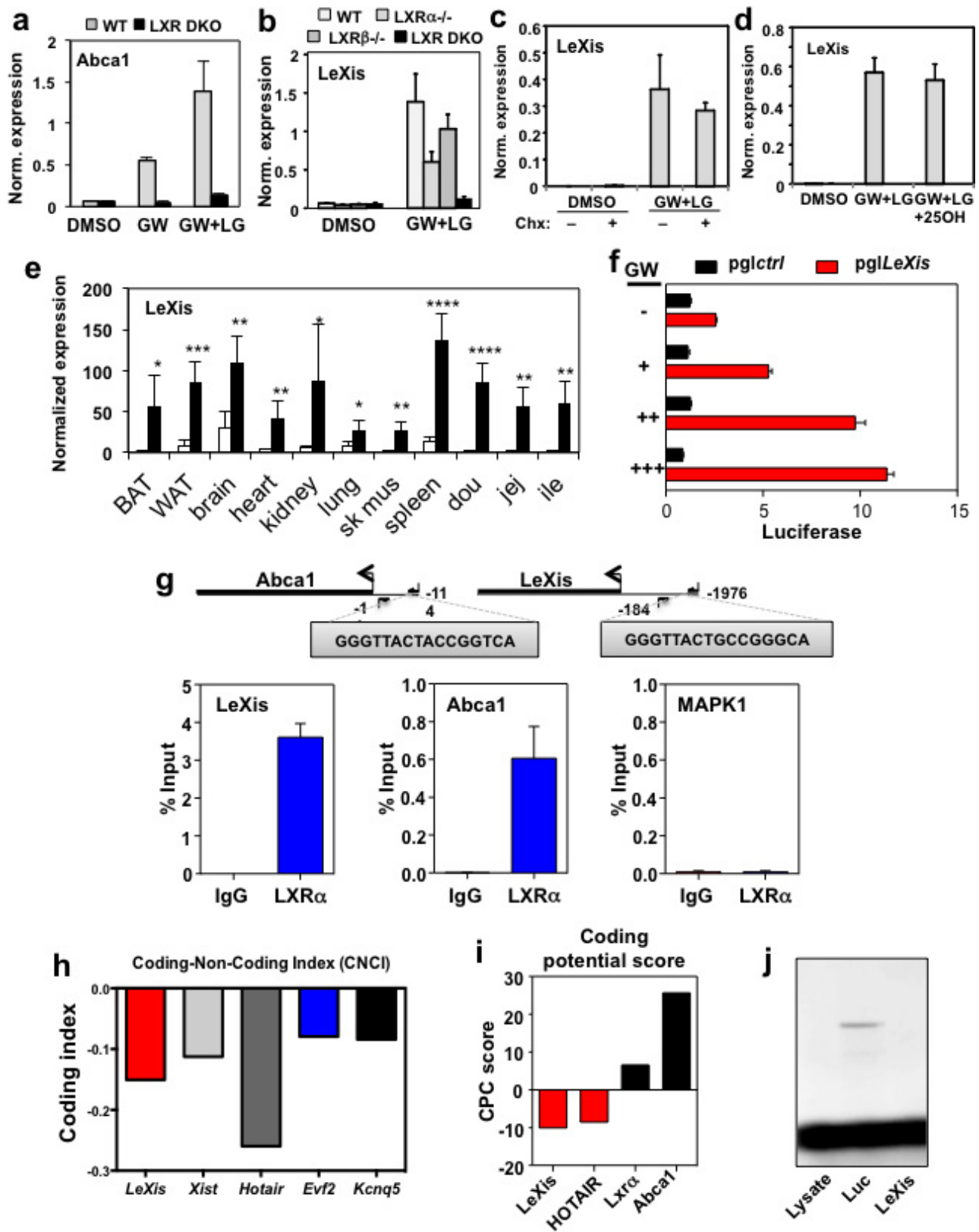
Extended Data Figure 1 | Identification of *LeXis* as an LXR-responsive lncRNA. **a**, qPCR analysis of gene expression in livers from mice gavaged with 40 mg kg⁻¹ GW3965 for 2 days. Mice were fasted for 4 h before collection ($n = 4$ per group). Values are mean \pm s.e.m. * $P < 0.05$; ** $P < 0.01$; *** $P < 0.001$ (unpaired two-tailed t -test). **b**, Volcano plot of RNA-seq results from primary hepatocytes treated for 16 h with 1 μ M GW3965. **c**, Relative expression of selected LXR target genes identified in

the RNA-seq study shown in **b**. Fold change represents ratio of transcript expression in GW3965 compared to DMSO treatment samples. Cut-off fold induction of 1.1 used (total 4,708 transcripts induced). **d**, Heat map representation of the results of transcriptional profiling (Agilent SurePrint G3 Gene Expression arrays) of primary hepatocytes treated with 1 μ M GW3965 for 16 h. Data were analysed using GeneSpring software.



Extended Data Figure 2 | Schematic of the *LeXis* gene locus and its RNA transcripts. **a**, UCSC genome browser view of RNA-seq transcriptional signatures at the *Abca1* and *LeXis* locus in mouse primary hepatocytes

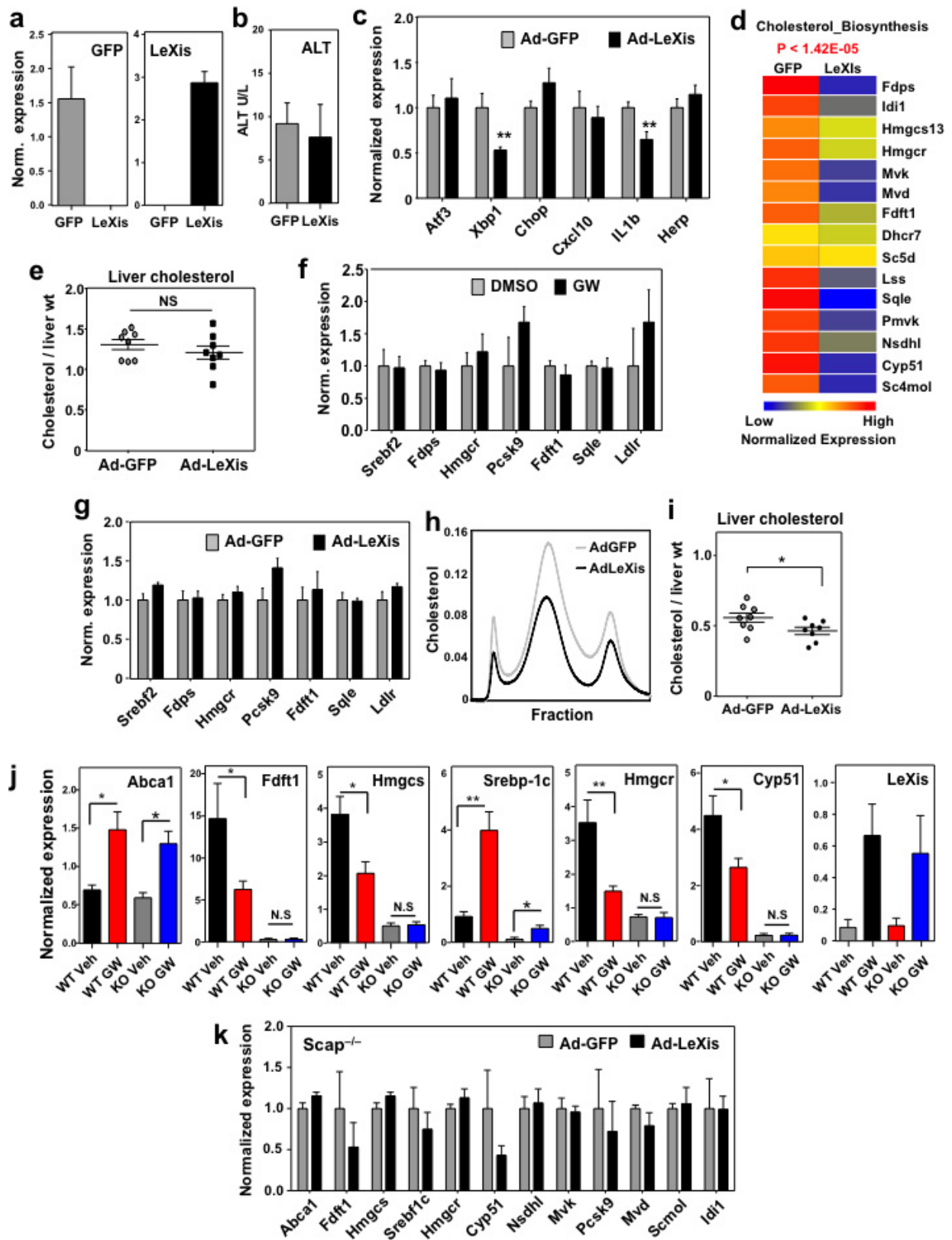
treated with 1 μ M GW3965 for 16 h. **b**, Exon structure of major and minor *LeXis* transcripts identified by RACE, aligned for comparison to existing annotation in the indicated databases.



Extended Data Figure 3 | See next page for caption.

Extended Data Figure 3 | Regulation of *LeXis* expression. **a**, qPCR analysis of primary mouse hepatocytes from wild-type or double knockout (*LXR α* ^{-/-} and *LXR β* ^{-/-}) mice treated with 1 μ M GW3965 and/or 50 nM LG268. Results are representative of four independent experiments. **b**, *LeXis* expression in primary mouse hepatocytes from wild-type, *LXR α* ^{-/-}, *LXR β* ^{-/-} or double knockout mice treated with GW3965 and LG268. Results are representative of three independent experiments. **c**, *LeXis* expression in primary hepatocytes treated with GW3965 and LG268 in the presence or absence of the protein synthesis inhibitor cycloheximide (Chx, 1 μ g μ l⁻¹). Results are representative of three independent experiments. **d**, *LeXis* expression in primary hepatocytes treated with GW3965 and LG268 (50 nM) in the presence or absence of 25-hydroxycholesterol (25OH, 2.5 μ M). Results are representative of three independent experiments. **e**, Gene expression in tissues from C57BL/6 mice gavaged with 40 mg kg⁻¹ GW3965 for 3 days ($n = 5$ per group). * $P < 0.05$; ** $P < 0.01$; *** $P < 0.001$; **** $P < 0.0001$ (unpaired two-tailed

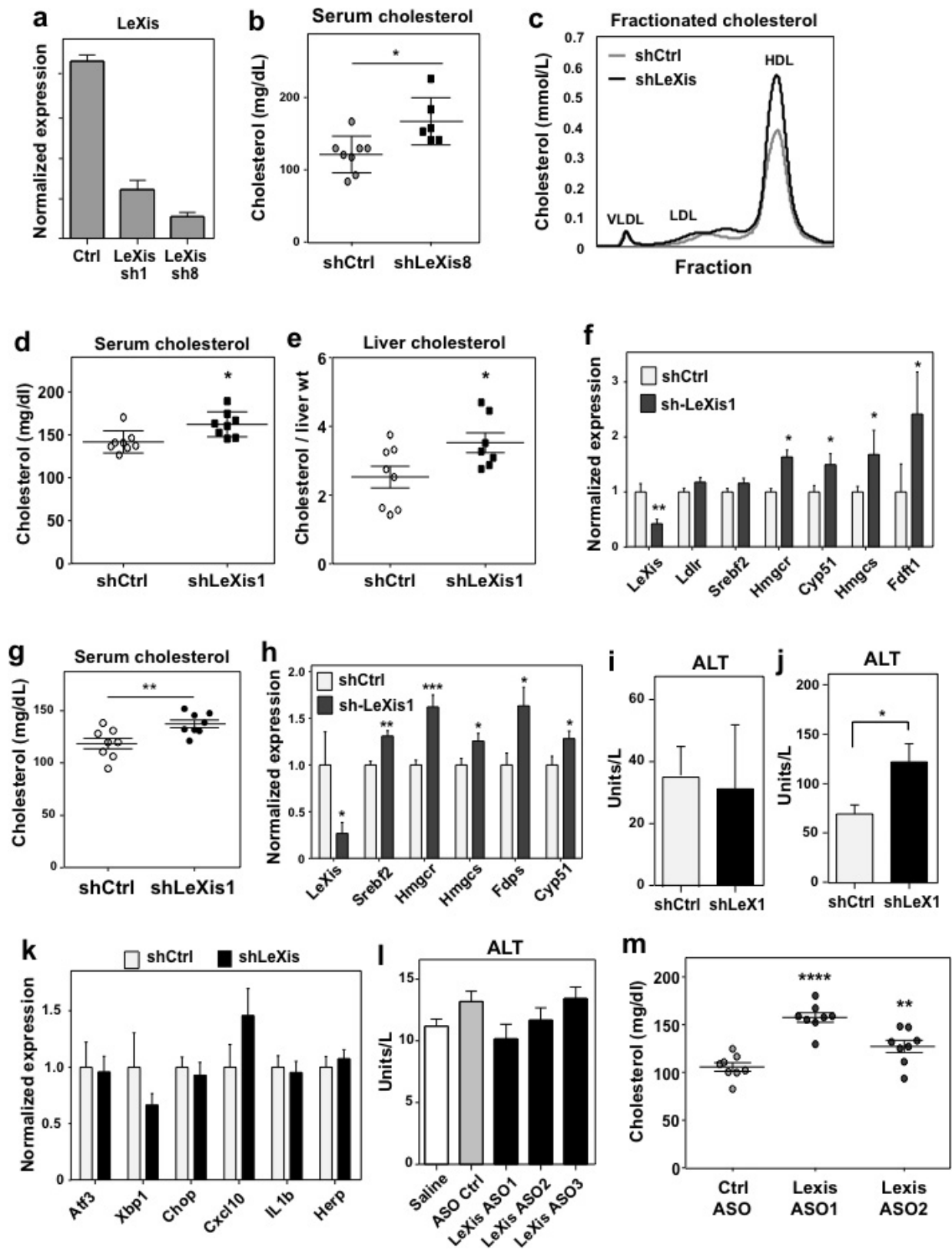
t -test). **f**, Relative firefly luciferase activity measured from the pgl4.10 vector or pgl4.10 with the *LeXis* promoter cloned upstream of luciferase. Reporters were co-transfected in HEK293 cells and treated with GW3965 for 24 h. Activity is normalized to *Renilla* luciferase internal control. **g**, Analysis of LXR α binding to the *LeXis* promoter in mouse liver by ChIP-qPCR. Schematic shows primer pair positions relative to the LXR-response element in the *LeXis* and *Abca1* (positive control) promoters. Primers flanking a region of the MAP kinase I promoter served as a negative control. ChIP values are presented as percentage of input DNA ($n = 4$ per group). Values are mean \pm s.e.m. (**e**, **g**) or mean \pm s.d. (**a–d**). **h**, Prediction of coding potential using the coding-non-coding index (CNCI) software. Negative value indicates low coding potential. **i**, Comparison of protein coding potential using coding potential calculator (CPC) score for *LeXis*, the non-coding gene *HOTAIR*, and control protein-coding transcripts. **j**, *In vitro* translation of *LeXis* and luciferase control RNAs.



Extended Data Figure 4 | See next page for caption.

Extended Data Figure 4 | *LeXis* modulates the expression of genes linked to sterol synthesis. **a**, Gene expression in livers obtained after 6 days of transduction with Ad-GFP or Ad-*LeXis* ($n = 8$ per group). **b**, Serum alanine aminotransferase activity in chow-fed mice transduced with Ad-GFP or Ad-*LeXis* for 6 days ($n = 8$ per group). **c**, Gene expression in livers obtained after 6 days of transduction with Ad-GFP or Ad-*LeXis* ($n = 8$ per group). **d**, Unbiased pathway analysis (GeneSpring software) of the results from transcriptional profiling of livers treated with Ad-GFP or Ad-*LeXis* ($n = 4$ per group). **e**, Hepatic cholesterol content normalized to liver mass in wild-type mice transduced with Ad-GFP or Ad-*LeXis* ($n = 8$ per group). **f**, Gene expression in mouse hepatocytes treated overnight with $1 \mu\text{M}$ GW3965. Results are representative of two independent

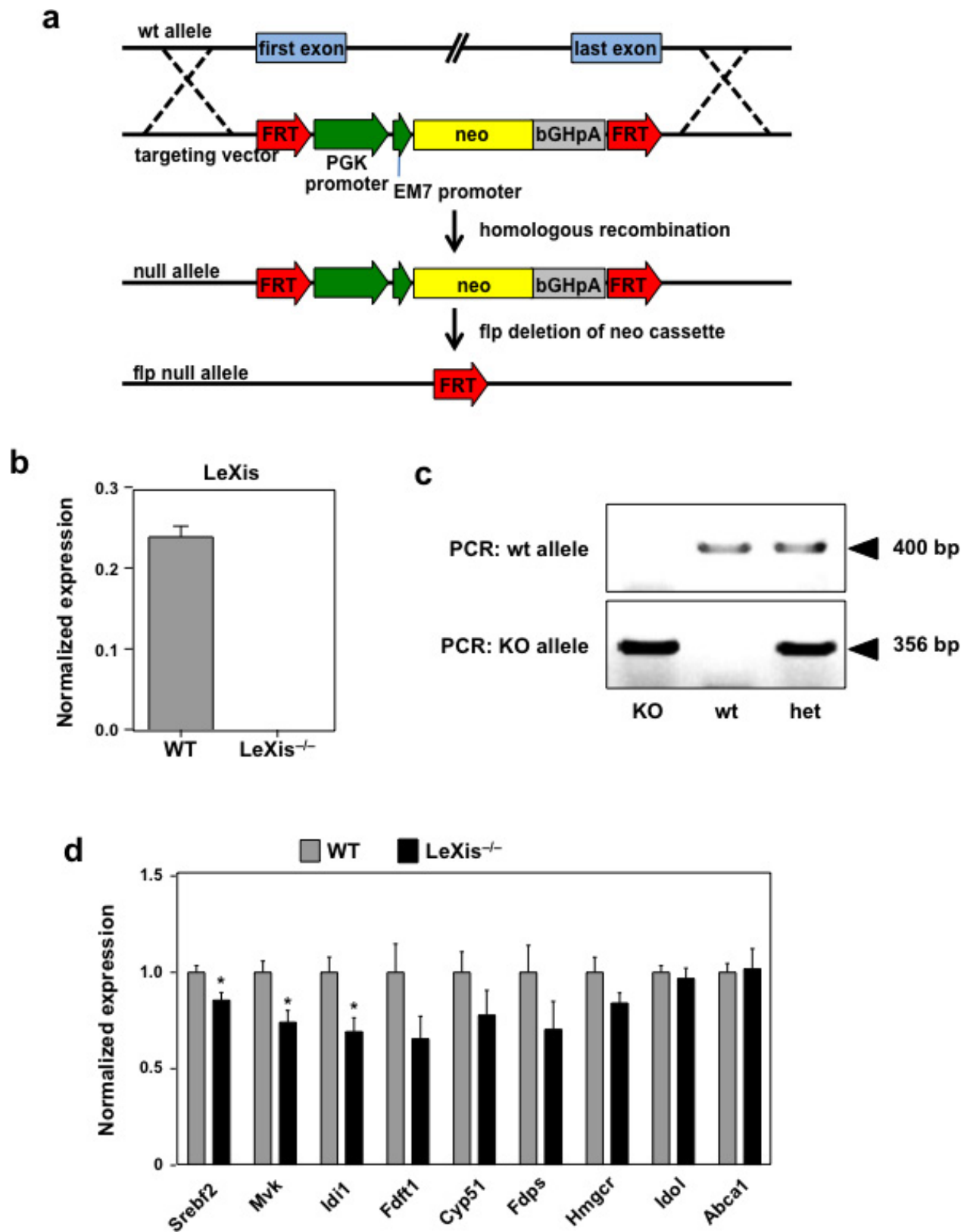
experiments. **g**, Gene expression in mouse hepatocytes treated overnight with Ad-GFP or Ad-*LeXis* for 24 h. Results are representative of two independent experiments. **h**, Cholesterol levels in pooled fractionated serum from *Ldlr*^{-/-} mice transduced with Ad-GFP or Ad-*LeXis*. **i**, Hepatic cholesterol content normalized to liver mass in *Ldlr*^{-/-} mice transduced with Ad-GFP or Ad-*LeXis* ($n = 8$ per group). **j**, Gene expression in livers from chow-fed wild-type or liver-specific *Scap*^{-/-} mice gavaged with 40 mg kg^{-1} GW3965 for 2 days ($n = 5$ (WT Veh), 8 (WT GW), 5 (KO Veh) and 7 (KO GW)). **k**, Gene expression in livers from *Scap*^{-/-} chow-fed mice transduced with Ad-GFP or Ad-*LeXis* for 6 days ($n = 5$ per group). Values are mean \pm s.e.m. (**a–c**, **e**, **i–k**) or mean \pm s.d. (**f**, **g**). * $P < 0.05$; ** $P < 0.01$ (unpaired two-tailed *t*-test).



Extended Data Figure 5 | See next page for caption.

Extended Data Figure 5 | Inhibition of *LeXis* expression alters serum cholesterol level. **a**, *In vitro* validation of *LeXis* knockdown using shLeXis1 and shLeXis8 vectors. Results are representative of three independent experiments. **b**, Total serum cholesterol measured in C57BL/6 mice fed 2 weeks of a Western diet and transduced with adenovirus shCtrl or shLeXis8 for 6 days ($n = 6-8$ per group). **c**, Cholesterol levels in pooled fractionated serum from mice transduced with shCtrl or shLeXis adenovirus. **d**, Total serum cholesterol from male C57BL/6 mice fed a Western diet for 2 weeks and then transduced with control (shCtrl) or adenoviral vectors expressing shRNA targeting *LeXis* (shLeXis1) ($n = 8$ per group). **e**, Hepatic cholesterol content normalized to liver mass for the mice shown in **d** ($n = 8$ (shCtrl) and 7 (shLeXis1)). **f**, Gene expression in livers of mice fed a Western diet for 2 weeks and then transduced with shCtrl or shLeXis ($n = 8$ (shCtrl) and 7 (shLeXis1)). **g**, Total plasma cholesterol levels in chow-fed C57BL/6 mice transduced with shCtrl or shLeXis adenovirus and gavaged with 40 mg kg^{-1}

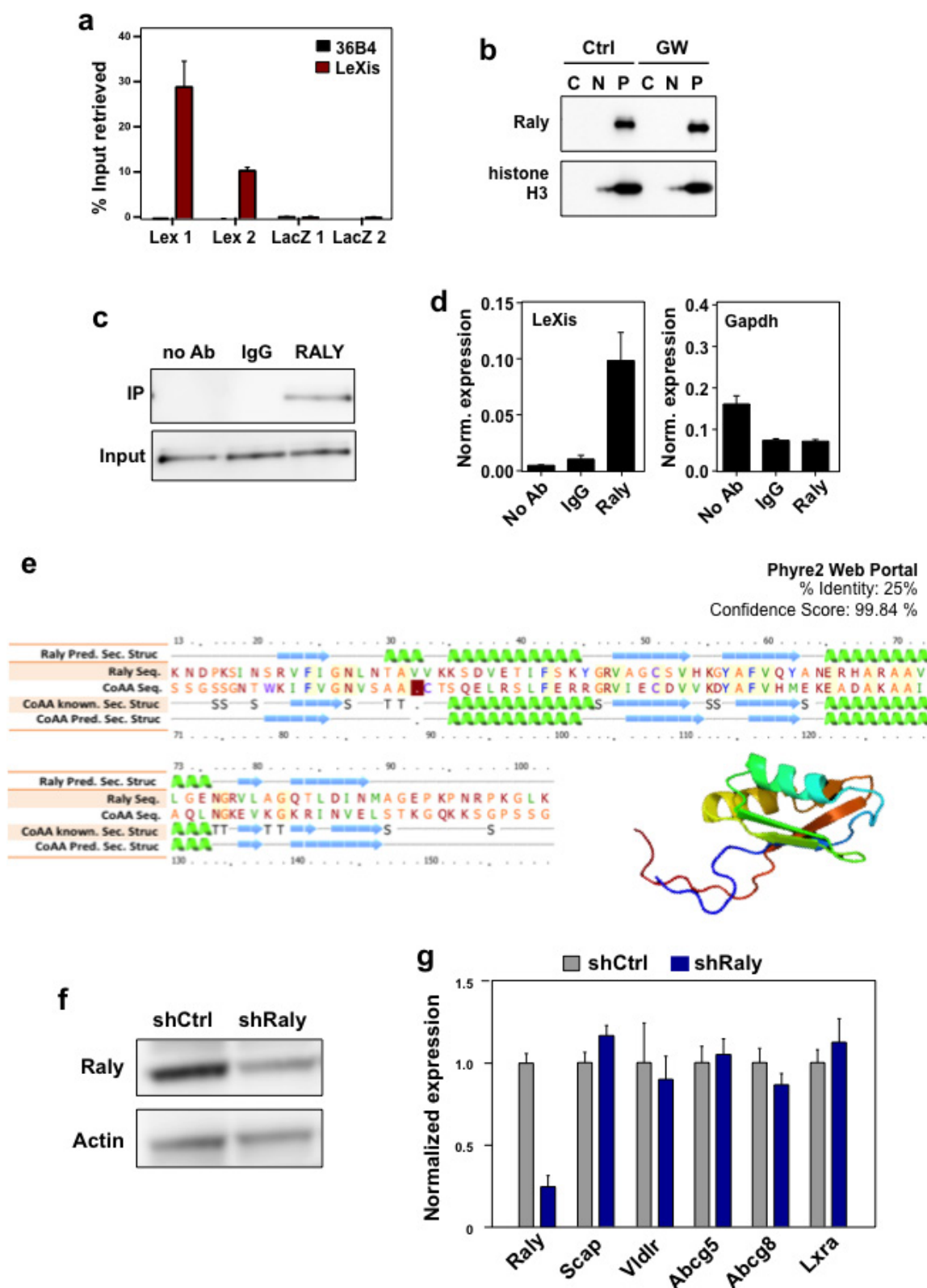
GW3965 for 6 days ($n = 8$ per group). **h**, Gene expression in livers of chow-fed C57BL/6 mice transduced with shCtrl or shLeXis adenovirus and gavaged with 40 mg kg^{-1} GW3965 for 6 days ($n = 8$ per group). **i**, Serum alanine aminotransferase activity from mice in **h**. **j**, Serum alanine aminotransferase activity from mice in **d**. **k**, Gene expression in livers of mice fed a Western diet for 2 weeks and then transduced with shCtrl or shLeXis ($n = 8$ (shCtrl) and 7 (shLeXis1)). **l**, Serum alanine aminotransferase activity from C57BL/6 mice on a chow diet administered 25 mg kg^{-1} ASOs intraperitoneally on days 1, 4 and 7, and gavaged with 40 mg kg^{-1} GW3965 on days 4, 7 and 8 ($n = 5$ per group). **m**, Total serum cholesterol from C57BL/6 mice on a chow diet administered 25 mg kg^{-1} ASOs intraperitoneally on days 1, 3 and 5, and gavaged with 40 mg kg^{-1} GW3965 on days 5 and 6 ($n = 8$ per group). Values are mean \pm s.d. (**a**) or mean \pm s.e.m. (**f**, **h-m**). * $P < 0.05$; ** $P < 0.01$; *** $P < 0.001$ (unpaired two-tailed *t*-test (**b**, **d-h**, **j**) and ANOVA with multi-group comparison (**m**)).



Extended Data Figure 6 | Generation of global *LeXis*^{-/-} mice.

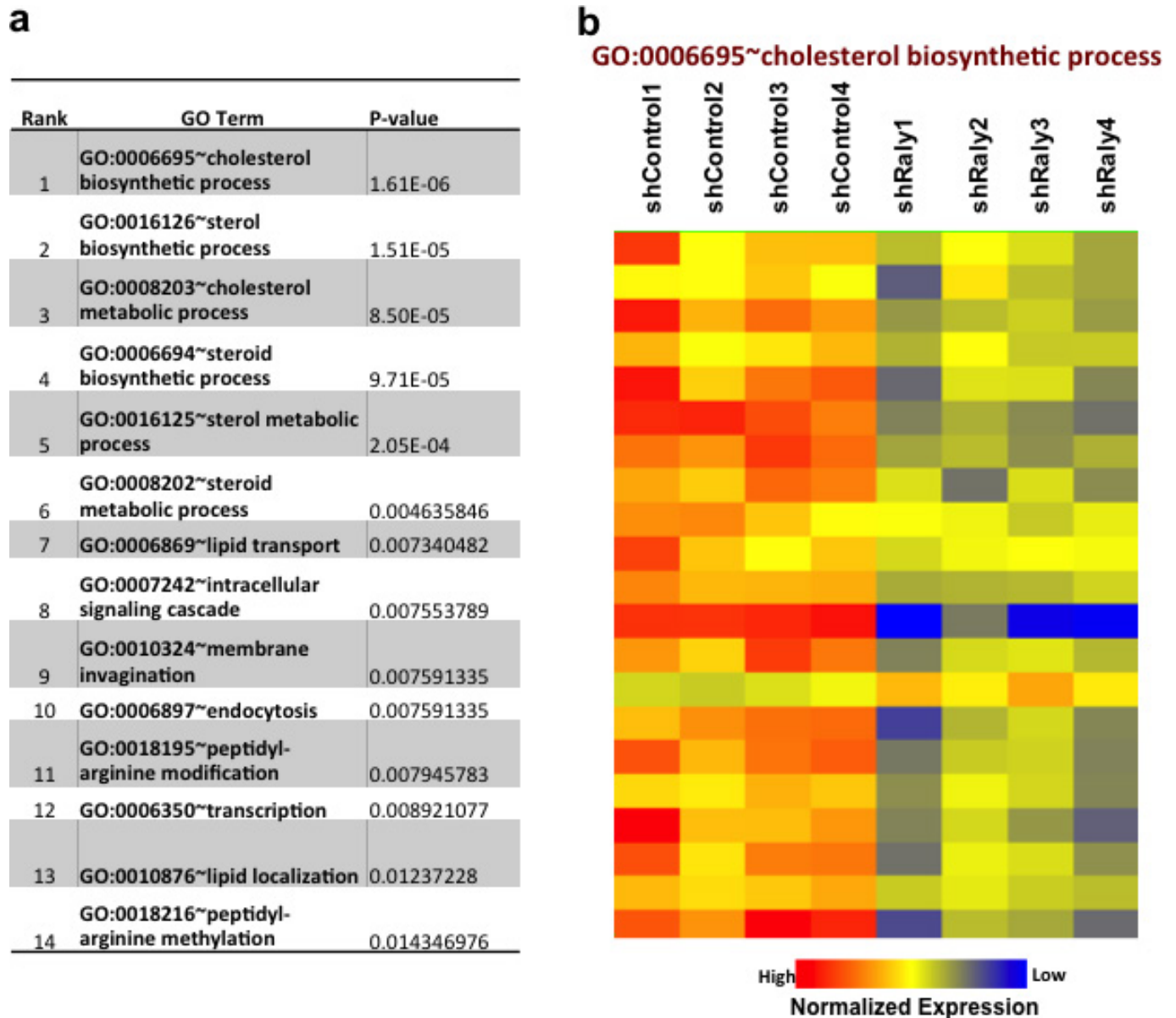
a, Schematic of knockout strategy. Vector construct designed to ablate entire *LeXis* transcript. Targeted mice were crossed with *Flp*^{-/-} (also known as *Hpd*^{-/-}) mice to excise the Neo cassette since it contains an active bi-directional promoter. **b**, **c**, Gene expression ($n = 3$ per group)

and PCR genotyping strategy for *LeXis*^{-/-} mice. **d**, Gene expression from C57BL/6 wild-type or *LeXis*^{-/-} mice fed on Western diet for 3 weeks ($n = 11$ (WT) and 7 (*LeXis*^{-/-})). All values are mean \pm s.e.m. * $P < 0.05$ (unpaired two-tailed *t*-test).

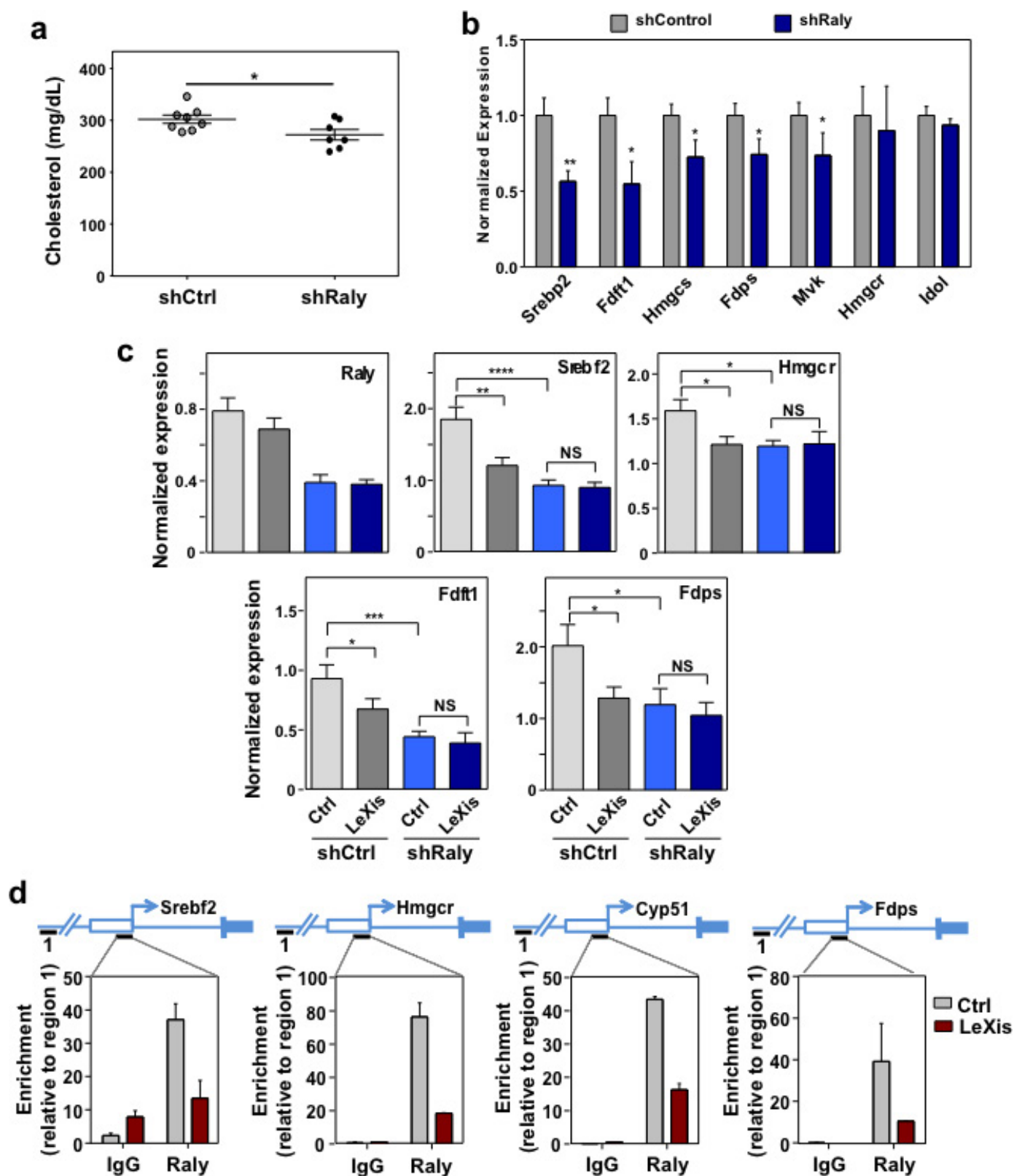


Extended Data Figure 7 | Identification of RALY as a LeXis-interacting protein. **a**, Complimentary biotin-labelled tiling oligonucleotides incubated with cellular extracts from liver. Probes sets designed to retrieve *LeXis* (Lex 1 and 2) or *LacZ* (LacZ 1 and 2). Percentage input of retrieved *LeXis* and *36B4* are shown ($n = 4$ per group). **b**, Cellular contents separated into cytoplasmic soluble (C), nuclear soluble (N) and insoluble (pellet, P) fractions were analysed by western blotting with anti-RALY and anti-histone H3 antibodies. **c**, Antibodies were incubated with cellular lysates from mouse hepatocytes and interaction with endogenous RALY was assessed after immunoprecipitation and western blot. **d**, Complexes from

b were analysed for presence of *LeXis* or *Gapdh* by reverse transcription qPCR (RT-qPCR) and signals were normalized to *36B4* ($n = 4$ per group). **e**, Sequence alignment, predicted secondary structure, and 3D model of RALY are shown as reported using the Phyre2 (Protein Homology/analogue Recognition Engine V 2.0) web portal. **f**, Western blot for RALY from livers transduced with adenoviral vectors expressing control shRNA (shCtrl) or *Raly* shRNA (shRaly) ($n =$ pooled 4 animals per group). **g**, Gene expression from liver from 14-week-old chow-fed male C57BL/6 mice transduced with control (shCtrl) or shRaly ($n = 8$ per group). Values are mean \pm s.d. (a) or mean \pm s.e.m. (g).

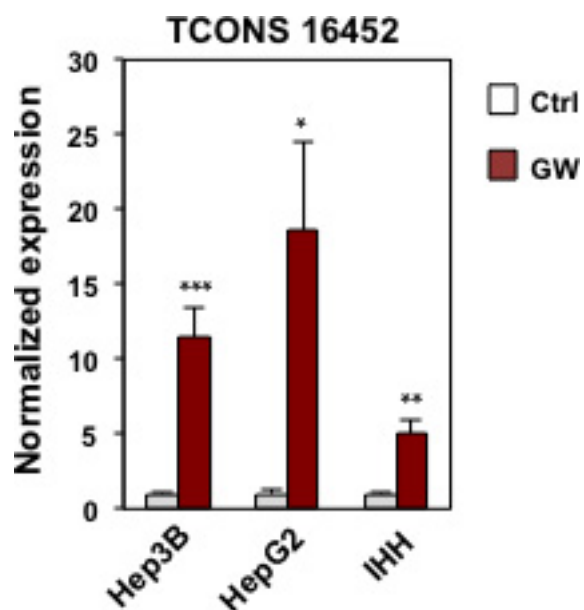


Extended Data Figure 8 | Knockdown of RALY preferentially affects pathways link to cholesterol metabolism in mouse liver. a, b, Most significant Gene Ontology terms from microarray analysis from livers treated with shCtrl or shRaly. Analysis performed using GeneSpring and DAVID.



Extended Data Figure 9 | RALY is required for *LeXis* mediated effects on cholesterologenesis. **a**, Total serum cholesterol levels in *Ldlr*^{-/-} mice transduced with shCtrl or shRaly for 6 days ($n=8$ (shCtrl) and 7 (shRaly)). **b**, Gene expression from liver obtained from *Ldlr*^{-/-} mice transduced with shCtrl or shRaly for 6 days ($n=8$ (shCtrl) and 7 (shRaly)). **c**, Gene expression from C57BL/6 mice transduced with control (Ad-GFP) or Ad-*LeXis* (1.0×10^9 p.f.u.) and shCtrl or shRaly (2.0×10^9 p.f.u.) ($n=7$

(ctrl/shCtrl and *LeXis*/shRaly) and 8 (*LeXis*/shCtrl and ctrl/shRaly)). **d**, Recruitment of RALY in promoter regions as determined by ChIP analysis in livers transduced with control (Ad-GFP) or Ad-*LeXis*. Data expressed as percentage input retrieved normalized to an upstream site (region 1) ($n=3$ per group). Values are mean \pm s.e.m. (**b**, **c**) or mean \pm s.d. (**d**). * $P < 0.05$; ** $P < 0.01$; *** $P < 0.001$; **** $P < 0.0001$ (unpaired two-tailed *t*-test (**a**, **b**) and ANOVA with multi-group comparison (**c**)).



Extended Data Figure 10 | Batch genome conversion between mouse and human at *LeXis* gene locus. Gene expression for putative human non-coding RNA TCONS_00016452 in hepatocyte cell lines treated with 1 μ M GW3965 ($n = 3$ per group). Values are mean \pm s.d. * $P < 0.05$; ** $P < 0.01$; *** $P < 0.001$ (unpaired two-tailed t -test).

Overcoming EGFR(T790M) and EGFR(C797S) resistance with mutant-selective allosteric inhibitors

Yong Jia¹, Cai-Hong Yun^{2,3†}, Eunyong Park^{2,3}, Dalia Ercan⁴, Mari Manuia¹, Jose Juarez¹, Chunxiao Xu⁴, Kevin Rhee⁴, Ting Chen⁴, Haikuo Zhang⁴, Sangeetha Palakurthi⁵, Jaebong Jang^{2,3}, Gerald Lelais¹, Michael DiDonato¹, Badry Bursulaya¹, Pierre-Yves Michellys¹, Robert Eppele¹, Thomas H. Marsilje¹, Matthew McNeill¹, Wenshuo Lu¹, Jennifer Harris¹, Steven Bender¹, Kwok-Kin Wong^{4,5}, Pasi A. Jänne^{4,5} & Michael J. Eck^{2,3}

The epidermal growth factor receptor (EGFR)-directed tyrosine kinase inhibitors (TKIs) gefitinib, erlotinib and afatinib are approved treatments for non-small cell lung cancers harbouring activating mutations in the EGFR kinase^{1,2}, but resistance arises rapidly, most frequently owing to the secondary T790M mutation within the ATP site of the receptor^{3,4}. Recently developed mutant-selective irreversible inhibitors are highly active against the T790M mutant^{5,6}, but their efficacy can be compromised by acquired mutation of C797, the cysteine residue with which they form a key covalent bond⁷. All current EGFR TKIs target the ATP-site of the kinase, highlighting the need for therapeutic agents with alternative mechanisms of action. Here we describe the rational discovery of EAI045, an allosteric inhibitor that targets selected drug-resistant EGFR mutants but spares the wild-type receptor. The crystal structure shows that the compound binds an allosteric site created by the displacement of the regulatory C-helix in an inactive conformation of the kinase. The compound inhibits L858R/T790M-mutant EGFR with low-nanomolar potency in biochemical assays. However, as a single agent it is not effective in blocking EGFR-driven proliferation in cells owing to differential potency on the two subunits of the dimeric receptor, which interact in an asymmetric manner in the active state⁸. We observe marked synergy of EAI045 with cetuximab, an antibody therapeutic that blocks EGFR dimerization^{9,10}, rendering the kinase uniformly susceptible to the allosteric agent. EAI045 in combination with cetuximab is effective in mouse models of lung cancer driven by EGFR(L858R/T790M) and by EGFR(L858R/T790M/C797S), a mutant that is resistant to all currently available EGFR TKIs. More generally, our findings illustrate the utility of purposefully targeting allosteric sites to obtain mutant-selective inhibitors.

Diverse activating mutations within the EGFR kinase domain give rise to a subset of non-small cell lung cancers (NSCLCs). The L858R point mutation and small in-frame deletions in the region encoded by exon 19 are the most common mutations, and are among a subset of oncogenic EGFR alterations that confer enhanced sensitivity to EGFR-directed TKIs^{11–13}. The dose-limiting toxicity of anilinoquinazoline TKIs such as erlotinib and gefitinib arises from inhibition of wild-type EGFR in the skin and GI tract, thus this enhanced sensitivity relative to wild-type EGFR creates a therapeutic window that allows effective treatment of patients whose tumours are driven by these mutations. The T790M resistance mutation closes this window, in part by increasing the affinity of the mutant receptor for ATP, which in turn diminishes the potency of these ATP-competitive inhibitors¹⁴. Mutant-selective irreversible inhibitors, including the tool compound WZ4002 (ref. 15) and the clinical compounds osimertinib (AZD9291)^{6,16} and rociletinib (CO-1686)⁵, are based on a pyrimidine scaffold, and also

incorporate a Michael acceptor group that forms a covalent bond with Cys797 at the edge of the ATP binding pocket. Because they bind irreversibly, these agents overcome the enhanced ATP affinity conferred by the T790M mutation. Compounds of this class are demonstrating significant efficacy against T790M mutant tumours in ongoing clinical trials^{17,18}, and osimertinib was recently approved by the US Food and Drug Administration for patients with EGFR T790M-positive NSCLC following progression on previous EGFR TKI therapy. However, laboratory studies and early clinical experience indicate that the efficacy of these agents can be compromised by mutation of Cys797, which thwarts formation of the potency-conferring covalent bond^{7,15,19}.

Reasoning that an allosteric inhibitor could also overcome the enhanced ATP affinity conferred by the T790M mutation, we screened an ~2.5 million compound library using purified EGFR(L858R/T790M) kinase. The biochemical screen was carried out using 1 μ M ATP, and active compounds were counter-screened at 1 mM ATP and against wild-type EGFR to identify those that were potentially non-ATP-competitive and mutant selective. Among the compounds identified in the screen, EGFR allosteric inhibitor-1 (EAI001, Fig. 1a) was of particular interest owing to its potency and selectivity for mutant EGFR (half maximal inhibitory concentration (IC₅₀) = 0.024 μ M for L858R/T790M at 1 mM ATP, IC₅₀ > 50 μ M for wild-type EGFR). Further characterization of the mutant-selectivity of EAI001 revealed modest potency against the isolated L858R and T790M mutants (0.75 μ M and 1.7 μ M, respectively, Extended Data Fig. 1a). Medicinal-chemistry-based optimization of this compound yielded EAI045 (Fig. 1a), a 3 nM inhibitor of the L858R/T790M mutant with ~1000-fold selectivity versus wild-type EGFR at 1 mM ATP (Table 1). Enzyme kinetic characterization confirmed that the mechanism of inhibition was not competitive with respect to ATP (Table 1, Extended Data Fig. 1b). Profiling of EAI045 against a panel of 250 protein kinases revealed pronounced selectivity; no other kinases were inhibited by more than 20% at 1 μ M EAI045 (Extended Data Table 1). Evaluation of EAI045 in a safety pharmacology assay panel revealed also excellent selectivity against non-kinase targets (Extended Data Table 2).

The crystal structure of EAI001 bound to T790M-mutant EGFR showed that the compound binds in an allosteric pocket that is created in part by the outward displacement of the C-helix in the inactive conformation of the kinase (Fig. 1b, c, Extended Data Table 3). The compound binds as a 'three-bladed propeller' with the aminothiazole moiety inserted between the mutant gatekeeper methionine and active site residue Lys745. The phenyl substituent extends into a hydrophobic cleft at the back of the pocket and is in contact with Leu777 and Phe856. Finally, the 1-oxoisindolyl group extends along the C-helix towards the solvent exposed exterior. The compound also forms a hydrogen bond with Asp855 in the DFG motif. In further support of

¹Genomics Institute of the Novartis Research Foundation, San Diego, California 92121, USA. ²Department of Cancer Biology, Dana-Farber Cancer Institute, Boston, Massachusetts 02215, USA.

³Department of Biological Chemistry and Molecular Pharmacology, Harvard Medical School, Boston, Massachusetts 02115, USA. ⁴Lowe Center for Thoracic Oncology, Dana-Farber Cancer Institute, Boston, Massachusetts 02215, USA. ⁵Belfer Center for Applied Cancer Science, Dana-Farber Cancer Institute, Boston, Massachusetts 02215, USA. [†]Present address: Peking University Institute of Systems Biomedicine and Department of Biophysics, Peking University Health Science Center, Beijing 100191, China.

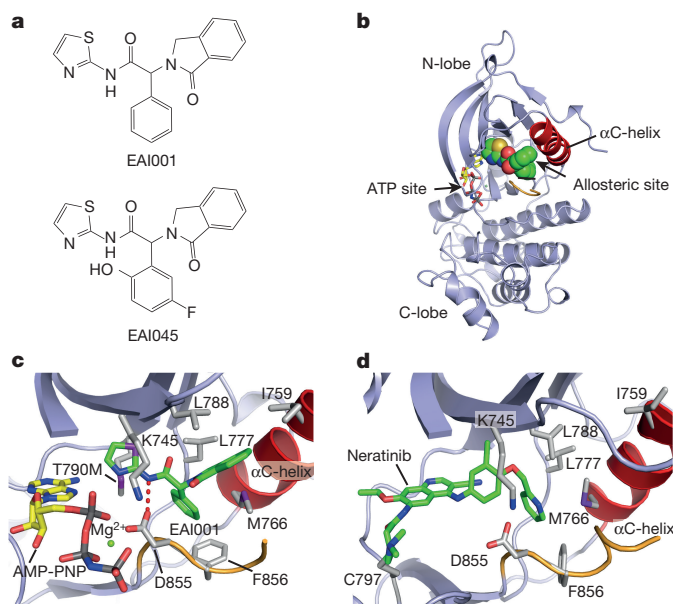


Figure 1 | Structure and binding mode of allosteric EGFR inhibitors.

a, Chemical structures of EAI001 and EAI045. **b**, Overall view of the structure of EGFR(T790M/V948R) mutant to EAI001 and AMP-PNP. EAI001 is shown in CPK-coloured form with carbon atoms in green. The V948R substitution was introduced to allow crystallization of the kinase in an inactive conformation⁸. **c**, Detailed view of the interactions of EAI001. A hydrogen bond with Asp855 in the DFG-motif of the kinase activation loop is shown as a dashed red line. **d**, The structure of irreversible inhibitor neratinib bound to EGFR(T790M) (PDB, 2JIV). Neratinib occupies the ATP site, but also extends into the allosteric pocket occupied by EAI001.

a non-ATP competitive mechanism, the ATP-analogue adenylyl-imidodiphosphate (AMP-PNP) is bound in the expected manner in the active site cleft (Fig. 1c).

Interestingly, the EGFR inhibitors neratinib²⁰ and lapatinib²¹ extend into the allosteric site and make interactions that resemble those of two of the three blades of the allosteric agents (Fig. 1d, Extended Data Fig. 2a, c). These ATP-competitive inhibitors are not mutant selective, and they span both the ATP and allosteric sites. Additionally, we note that the EGFR allosteric pocket is roughly analogous to a site in MEK1 that is targeted by a number of allosteric inhibitors that are now approved or in clinical trials²². Despite the similar location of the MEK allosteric site, there is no structural correspondence in the binding modes of the respective allosteric inhibitors (Extended Data Fig. 2a, b).

The mutant-specificity of the EGFR allosteric inhibitors arises from at least two effects. Most apparently, the direct contact of the aminothiazole group with the mutant gatekeeper methionine residue can explain the selectivity for the T790M mutant. Second, the compound cannot bind the fully inactive conformation of the wild-type kinase; simple modelling reveals steric clashes of EAI001 with Leu858 and Leu861 in the N-terminal portion of the activation loop (Extended Data Fig. 3). The L858R mutation rearranges this portion of the activation loop²³, thereby enlarging the allosteric pocket. EAI045 may also inhibit other mutants with a similar mechanism of activation, such as L861Q, but we do not expect it to inhibit most exon 19 deletion variants. These mutations shorten the loop leading into the C-helix and may therefore prevent opening of the allosteric pocket.

Initial studies of the cellular activity of EAI045 showed that it potently decreased, but did not completely eliminate, EGFR autophosphorylation in H1975 cells, an L858R/T790M-mutant NSCLC cell line (Fig. 2a). A similar effect was observed in NIH-3T3 cells stably transfected with the L858R/T790M mutant (Extended Data Fig. 4a). This inhibition was selective for mutant EGFR; EAI045 potently inhibited EGFR Y1173 phosphorylation in H1975 cells (half maximal effective concentration (EC_{50}) = 2 nM), but not in HaCaT cells, a keratinocyte

Table 1 | Inhibitory activity of EAI045 on wild type EGFR and selected mutants

ATP (μ M)	EAI045 IC_{50} (μ M)			
	Wild type	L858R	T790M	L858R/T790M
1	1.6	0.076	0.049	0.002
10	1.9	0.019	0.19	0.002
100	3.5	0.009	0.5	0.003
1000	4.3	0.009	0.6	0.003

cell line with wild-type EGFR (Extended Data Table 4). We observed an intermediate level of activity in the L858R-mutant H3255 cells, a pattern consistent with our biochemical inhibition data (Extended Data Table 4). Despite potent inhibition of mutant EGFR, EAI045 showed no anti-proliferative effect in the H1975 and H3255 cell lines with concentrations as high as 10 μ M (Extended Data Table 4). Profiling in a panel of EGFR-mutant Ba/F3 cells revealed that EAI045 inhibited proliferation of L858R/T790M and L858R mutant cells, but not the exon19del/T790M or parental Ba/F3 cells, indicative of on-target mutant-selective activity of the allosteric inhibitor (Extended Data Fig. 4b–e). However, half-maximal inhibition required \sim 10 μ M EAI045, a concentration much higher than the biochemical IC_{50} of the compound.

In light of the incomplete inhibition of EGFR autophosphorylation and the allosteric mechanism of action of EAI045, we wondered to what extent ligand stimulation would affect inhibition of the mutant receptor. We compared inhibition of EGFR Y1173 phosphorylation in H1975 cells in the presence and absence of exogenous EGF (10 ng ml⁻¹) using an ELISA-based assay. EAI045 inhibited EGFR phosphorylation with a similar EC_{50} irrespective of EGF stimulation, but notably, inhibition plateaued at 50% in the presence of ligand (Fig. 2b). This phenomenon suggests two populations of receptor, one that remains sensitive to the allosteric inhibitor upon ligand stimulation, and another, equal in number, that is rendered insensitive. Ligand-induced dimerization of the EGF receptor is known to induce an asymmetric interaction of the kinase domains⁸, and is an apparent potential source of two receptor populations with differential inhibitor sensitivity.

In the EGFR asymmetric dimer, the C-lobe of the ‘activator’ subunit impinges on the N-lobe of the ‘receiver’ subunit, inducing an active conformation in the receiver by reorienting the regulatory C-helix to its inward position (Fig. 2c). In wild-type EGFR, only the receiver subunit is activated. By contrast, both subunits in a mutant receptor are expected to be catalytically active, because oncogenic kinase domain mutations induce the active conformation even in the absence of ligand. As explained above, EAI045 binds a ‘C-helix out’ conformation of the kinase. In the receiver subunit but not the activator, outward displacement of the C-helix is impeded by the asymmetric dimer interaction. Therefore, we hypothesized that EAI045 was a potent inhibitor of the activator subunit of the mutant receptor, but a much less potent inhibitor of the receiver subunit, in which the C-helix is captive. Because the mutant receptor favours dimer formation^{24,25}, this effect could explain both the incomplete inhibition of EGFR autophosphorylation and the apparent disconnect in the biochemical and cellular potencies of the allosteric inhibitor. To test this notion, we exploited an I941R point mutation in the C-lobe of the kinase, which is known to block the asymmetric dimer interaction^{8,26}. The activity of the L858R/T790M mutant is dimerization-independent²⁶ and, as expected, transduction of Ba/F3 cells with EGFR(L858R/T790M/I941R) led to factor-independent proliferation. In support of our hypothesis, Ba/F3 cells bearing this dimerization-defective mutant were markedly more sensitive to the allosteric inhibitor (Fig. 2d).

The therapeutic antibody cetuximab targets the extracellular portion of the EGF receptor, blocking ligand binding and preventing dimer formation^{9,10}. The antibody is not effective clinically in EGFR-mutant NSCLC, and in cell-based studies cetuximab alone does not inhibit L858R/T790M or exon19del/T790M mutant EGFR, because their activity is independent of dimerization²⁶. However, we reasoned that

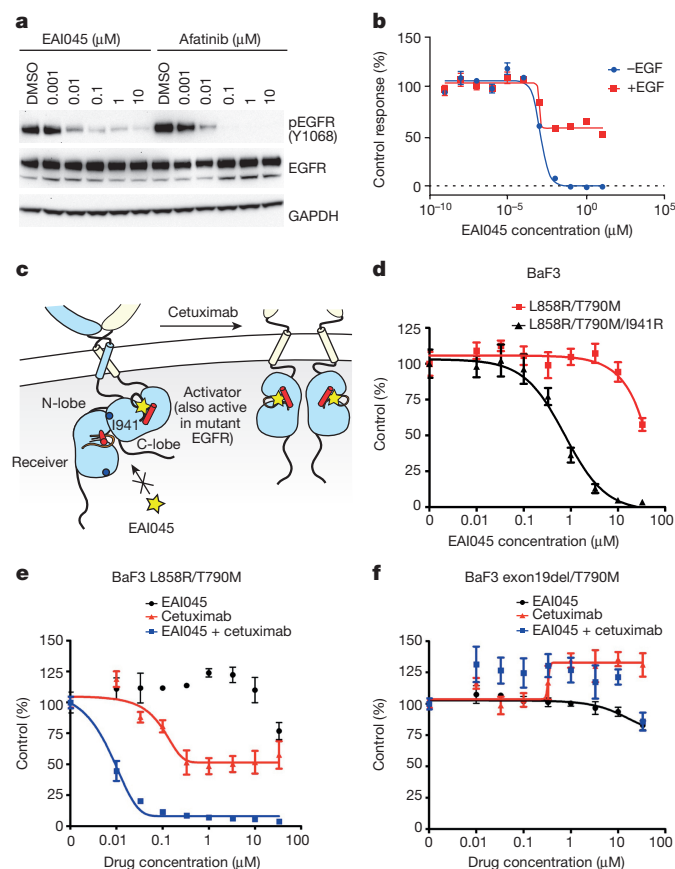


Figure 2 | Cellular activity and mechanism of synergy of EAI045 with cetuximab. **a**, Analysis of EAI045 inhibition of EGFR phosphorylation in H1975 cells by western blotting (anti-pY1068). A dose response study is shown at 3 h after compound addition for EAI045 and the irreversible quinazoline inhibitor afatinib (control). For gel source data, see Supplementary Fig. 1. **b**, The effect of EAI045 on EGFR target modulation in H1975 cells in the presence and absence of EGF. EGFR phosphorylation (pY1173) was measured using an ELISA-based assay; error bars indicate s.d. ($n = 3$). **c**, The allosteric pocket is differentially accessible in the two subunits of the asymmetric dimer. Unlike wild-type EGFR in which only the receiver subunit is active, both subunits are catalytically active in the L858R/T790M mutant. The activator subunit is more readily inhibited by allosteric agents (yellow star), because the C-helix can be readily displaced. By contrast, opening the allosteric pocket in the receiver subunit requires perturbing the dimer. Thus mutations that disrupt the asymmetric dimer (such as I941R, blue circle) or antibodies that block dimerization (cetuximab) should enhance the potency of allosteric agents. **d**, Inhibition of proliferation of Ba/F3 cells expressing L858R/T790M and L858R/T790M/I941R by EAI045. Addition of the dimer-disrupting I941R mutation markedly increased inhibition by EAI045. **e**, **f**, Treatment of EGFR-mutant Ba/F3 cells with EAI045 alone, in combination with cetuximab ($10 \mu\text{g ml}^{-1}$), or with cetuximab alone. Note the pronounced synergy with cetuximab that is observed only in the L858R/T790M model. The mean \pm s.d. ($n = 6$) is plotted for each drug and concentration (**d–f**).

cetuximab should synergize with a kinase-targeted allosteric inhibitor, by converting the inhibitor-resistant receiver population into a monomeric form that is remarkably sensitive to EAI045. Notably, in the presence of cetuximab ($10 \mu\text{g ml}^{-1}$), EAI045 inhibited proliferation of EGFR(L858R/T790M) Ba/F3 cells with an IC_{50} of approximately 10 nM, similar to its potency against this mutant in biochemical assays (Fig. 2e). In support of an on-target, mutant-selective effect of the allosteric agent, proliferation of Ba/F3 cells bearing EGFR(exon19del/T790M) was not inhibited by this combination (Fig. 2f).

We next tested the *in vivo* efficacy of EAI045 in genetically engineered mouse model of L858R/T790M-mutant-driven lung cancer²⁷, both alone and in combination with cetuximab. Mouse

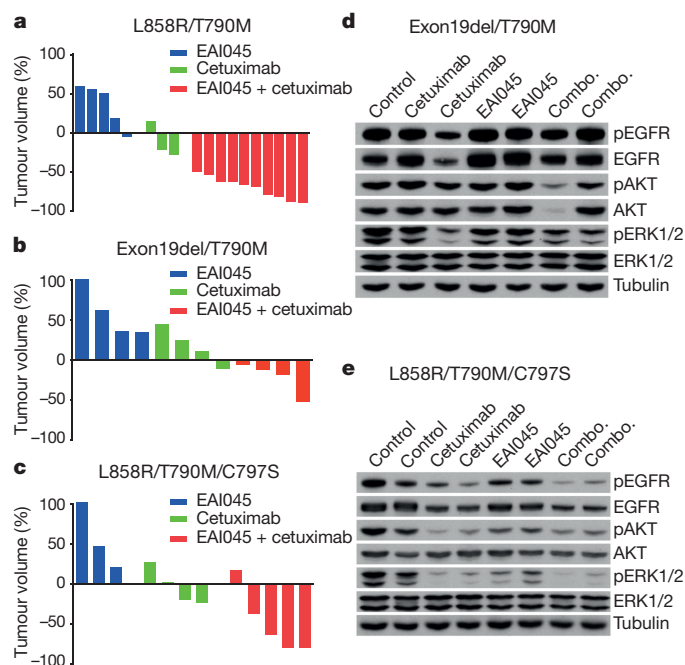


Figure 3 | EAI045 in combination with cetuximab induces tumour regression in genetically engineered mouse models of EGFR-mutant lung cancer. **a**, Mice bearing L858R/T790M mutant tumours were treated with EAI045 alone ($n = 5$), cetuximab alone ($n = 3$) or both agents in combination ($n = 10$). Tumour volumes were measured using MRI 4 weeks after initiation of treatment and are plotted for each animal in a ‘waterfall’ format. **b**, As in **a**, but in mice bearing exon19del/T790M mutant tumours ($n = 4, 4$ and 4). **c**, As in **a**, but in mice bearing L858R/T790M/C797S mutant tumours ($n = 3, 4$, and 5). **d**, **e**, Pharmacodynamic studies in exon19del/T790M and L858R/T790M/C797S mice. Tumour nodules from mice treated with EAI045 or cetuximab alone or with the combination (combo.) were analysed by western blotting with the indicated antibodies to examine the effect of treatment on EGFR signalling. Multiple independent mouse tumours were obtained and analysed, two independent and representative samples are shown. For gel source data, see Supplementary Fig. 1. Source data for tumour volume measurements are provided in Supplementary Fig. 2.

pharmacokinetic studies with EAI045 revealed a maximal plasma concentration of $0.57 \mu\text{M}$, a half-life of 2.15 h, and oral bioavailability of 26% after dosing at 20 mg kg^{-1} . In a 4-week efficacy study, mice were treated with EAI045 at 60 mg kg^{-1} by oral gavage once daily, either alone or together with cetuximab (1 mg intraperitoneally every other day). We observed marked tumour regressions in the L858R/T790M-mutant mice treated with the combination, whereas those treated with EAI045 alone did not respond (Fig. 3a). Cetuximab alone had a very modest effect in these mice, as previously observed²⁶. Mice bearing EGFR(exon19del/T790M) were treated using the same protocol, but as expected failed to respond to the combination therapy (Fig. 3b). Magnetic resonance imaging (MRI) studies of cohorts of L858R/T790M and exon19del/T790M mice after combination treatment for 1 or 2 weeks are shown in Extended Data Fig. 5.

Mutation of C797 is expected to confer resistance to all third-generation irreversible EGFR inhibitors that are active on the T790M-mutant EGFR, and a preliminary study reported the C797S alteration in 15 out of 67 patients (22%) with acquired resistance to AZD9291 (ref. 28). Mutations in C797 should not affect the efficacy of EAI045, as this residue is remote from the allosteric binding pocket. Consistent with this expectation, EAI045 in combination with cetuximab potently inhibited L858R/T790M/C797S Ba/F3 cells (Extended Data Fig. 5a) and treatment of genetically engineered L858R/T790M/C797S mice with EAI045 and cetuximab induced marked tumour shrinkage, similar to that observed in the L858R/T790M models (Fig. 3c, Extended Data Fig. 5b). Pharmacodynamic studies performed following two doses

of treatment demonstrated that EAI045 in combination with cetuximab effectively inhibited phosphorylation of EGFR and downstream signalling proteins in these mice, but not in mice bearing the insensitive exon19del/T790M mutation (Fig. 3d, e).

The compounds we describe here are among the first allosteric TKIs, and to our knowledge, the first targeting any receptor tyrosine kinase in a mutant-selective manner. Further study is required, but our findings suggest that EAI045 or a related compound in combination with an EGFR dimer-disrupting antibody such as cetuximab would be an effective strategy for treating L858R/T790M-mutant-driven lung cancers, as well as those driven by the triple L858R/T790M/C797S mutation, which are resistant to all current EGFR-targeted therapies. EAI045 and cetuximab exhibit mechanistic synergy, a valuable property for combination agents because it lowers the dose required for efficacy. Ideally, chemotherapeutic agents used in combination should also have non-overlapping mechanisms of toxicity and sensitivity to resistance mutations. EAI045 meets these criteria as well; its lack of activity on wild-type EGFR and other kinases suggest that its dose-limiting toxicity is unlikely to be related to that of cetuximab and ATP-competitive EGFR inhibitors. In addition, given its distinct binding site, its sensitivity to resistance-conferring mutations is expected to be divergent from that of both cetuximab and ATP-site inhibitors. For these reasons, we speculate that an allosteric agent like EAI045 could be used in combination with ATP-site-directed inhibitors, with the goal of preventing the emergence of treatment-associated resistance mutations in the receptor itself.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 15 October 2015; accepted 29 March 2016.

Published online 25 May 2016.

- Mok, T. S. *et al.* Gefitinib or carboplatin-paclitaxel in pulmonary adenocarcinoma. *N. Engl. J. Med.* **361**, 947–957 (2009).
- Yu, H. A. & Pao, W. Targeted therapies: Afatinib—new therapy option for EGFR-mutant lung cancer. *Nat. Rev. Clin. Oncol.* **10**, 551–552 (2013).
- Gainor, J. F. & Shaw, A. T. Emerging paradigms in the development of resistance to tyrosine kinase inhibitors in lung cancer. *J. Clin. Oncol.* **31**, 3987–3996 (2013).
- Chong, C. R. & Jänne, P. A. The quest to overcome resistance to EGFR-targeted therapies in cancer. *Nat. Med.* **19**, 1389–1400 (2013).
- Walter, A. O. *et al.* Discovery of a mutant-selective covalent inhibitor of EGFR that overcomes T790M-mediated resistance in NSCLC. *Cancer Discov.* **3**, 1404–1415 (2013).
- Finlay, M. R. *et al.* Discovery of a potent and selective EGFR inhibitor (AZD9291) of both sensitizing and T790M resistance mutations that spares the wild type form of the receptor. *J. Med. Chem.* **57**, 8249–8267 (2014).
- Thress, K. S. *et al.* Acquired EGFR C797S mutation mediates resistance to AZD9291 in non-small cell lung cancer harboring EGFR T790M. *Nat. Med.* **21**, 560–562 (2015).
- Zhang, X., Gureasko, J., Shen, K., Cole, P. A. & Kuriyan, J. An allosteric mechanism for activation of the kinase domain of epidermal growth factor receptor. *Cell* **125**, 1137–1149 (2006).
- Goldstein, N. I., Prewett, M., Zuklys, K., Rockwell, P. & Mendelsohn, J. Biological efficacy of a chimeric antibody to the epidermal growth factor receptor in a human tumor xenograft model. *Clin. Cancer Res.* **1**, 1311–1318 (1995).
- Li, S. *et al.* Structural basis for inhibition of the epidermal growth factor receptor by cetuximab. *Cancer Cell* **7**, 301–311 (2005).
- Paez, J. G. *et al.* EGFR mutations in lung cancer: correlation with clinical response to gefitinib therapy. *Science* **304**, 1497–1500 (2004).
- Pao, W. *et al.* EGF receptor gene mutations are common in lung cancers from “never smokers” and are associated with sensitivity of tumors to gefitinib and erlotinib. *Proc. Natl Acad. Sci. USA* **101**, 13306–13311 (2004).
- Lynch, T. J. *et al.* Activating mutations in the epidermal growth factor receptor underlying responsiveness of non-small-cell lung cancer to gefitinib. *N. Engl. J. Med.* **350**, 2129–2139 (2004).
- Yun, C. H. *et al.* The T790M mutation in EGFR kinase causes drug resistance by increasing the affinity for ATP. *Proc. Natl Acad. Sci. USA* **105**, 2070–2075 (2008).
- Zhou, W. *et al.* Novel mutant-selective EGFR kinase inhibitors against EGFR T790M. *Nature* **462**, 1070–1074 (2009).
- Cross, D. A. *et al.* AZD9291, an irreversible EGFR TKI, overcomes T790M-mediated resistance to EGFR inhibitors in lung cancer. *Cancer Discov.* **4**, 1046–1061 (2014).
- Sequist, L. V. *et al.* Rociletinib in EGFR-mutated non-small-cell lung cancer. *N. Engl. J. Med.* **372**, 1700–1709 (2015).
- Jänne, P. A. *et al.* AZD9291 in EGFR inhibitor-resistant non-small-cell lung cancer. *N. Engl. J. Med.* **372**, 1689–1699 (2015).
- Ercan, D. *et al.* EGFR mutations and resistance to irreversible pyrimidine-based EGFR inhibitors. *Clin. Cancer Res.* **21**, 3913–3923 (2015).
- Tsou, H. R. *et al.* Optimization of 6,7-disubstituted-4-(arylamino)quinoline-3-carbonitriles as orally active, irreversible inhibitors of human epidermal growth factor receptor-2 kinase activity. *J. Med. Chem.* **48**, 1107–1131 (2005).
- Wood, E. R. *et al.* A unique structure for epidermal growth factor receptor bound to GW572016 (Lapatinib): relationships among protein conformation, inhibitor off-rate, and receptor activity in tumor cells. *Cancer Res.* **64**, 6652–6659 (2004).
- Zhao, Y. & Adjei, A. A. The clinical development of MEK inhibitors. *Nat. Rev. Clin. Oncol.* **11**, 385–400 (2014).
- Yun, C. H. *et al.* Structures of lung cancer-derived EGFR mutants and inhibitor complexes: mechanism of activation and insights into differential inhibitor sensitivity. *Cancer Cell* **11**, 217–227 (2007).
- Red Brewer, M. *et al.* Mechanism for activation of mutated epidermal growth factor receptors in lung cancer. *Proc. Natl Acad. Sci. USA* **110**, E3595–E3604 (2013).
- Shan, Y. *et al.* Oncogenic mutations counteract intrinsic disorder in the EGFR kinase and promote receptor dimerization. *Cell* **149**, 860–870 (2012).
- Cho, J. *et al.* Cetuximab response of lung cancer-derived EGF receptor mutants is associated with asymmetric dimerization. *Cancer Res.* **73**, 6770–6779 (2013).
- Li, D. *et al.* Bronchial and peripheral murine lung carcinomas induced by T790M-L858R mutant EGFR respond to HKI-272 and rapamycin combination therapy. *Cancer Cell* **12**, 81–93 (2007).
- Oxnard, G. R. *et al.* in *16th World Conference on Lung Cancer* (Denver, Colorado, 2015).

Supplementary Information is available in the online version of the paper.

Acknowledgements This work was supported in part by NIH grants CA116020 (M.J.E.), CA154303 (M.J.E., K.-K.W. and P.A.J.), CA120964 (K.-K.W.) and CA135257 (P.A.J.), and by the Gross-Loh Family Fund for Lung Cancer Research (K.-K.W.). We thank N. Gray for helpful comments on the manuscript.

Author Contributions M.J.E., P.A.J., K.-K.W., Y.J., G.L., P.-Y.M., J.H., and S.B. coordinated the study. Y.J., M.M., J. Juarez, M.D., B.B., E.P., C.-H.Y., D.E., C.X., K.R., T.C., H.Z., S.P., and J. Jang designed and performed experiments. Y.J., M.M., J. Juarez, M.D., B.B., S.B., E.P., C.-H.Y., D.E., C.X., K.R., M.J.E., P.J., and K.-K.W. interpreted data. M.M., J. Juarez, M.D., G.L., P.-Y.M., R.E., T.H.M., M.M., C.-H.Y. and W.L. prepared reagents. Y.J., K.-K.W., P.A.J. and M.J.E. wrote and edited the manuscript.

Author Information The crystal structure of EGFR(T790M/V948R) in complex with EAI001 has been deposited in the Protein Data Bank under accession number 5D41. Reprints and permissions information is available at www.nature.com/reprints. The authors declare competing financial interests: details are available in the online version of the paper. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to M.J.E. (eck@crystal.harvard.edu).

METHODS

EGFR protein expression and purification. Constructs spanning residues 696–1022 of the human EGFR (including wild type, L858R, L858R/T790M, T790M, and T790M/V948R mutant sequences) were prepared in a GST-fusion format using the pTriEX system (Novagen) for expression in Sf9 insect cells essentially as described^{14,23}. EGFR kinase proteins were purified by glutathione-affinity chromatography followed by size-exclusion chromatography after cleavage with Tomato etch virus (TEV) or thrombin to remove the GST fusion partner following established procedures^{14,23}.

High-throughput screening. Purified EGFR(L858R/T790M) enzyme was screened against Novartis compound collection of ~2.5 million using homogeneous time-resolved fluorescence (HTRF)-based biochemical assay format. The screening was performed at 1 μ M ATP using a single compound concentration (12.5 μ M). 1,322 top hits were picked for follow-up IC₅₀ confirmation. IC₅₀ values were determined at both 1 μ M and 1 mM ATP to identify both ATP competitive and non-competitive compounds. Hits were also counter-screened against wild-type EGFR to evaluate the mutant selectivity.

HTRF-based EGFR biochemical assays. Biochemical assays for wild-type EGFR and each mutant were carried out using a HTRF assay as described previously²⁹. Assays were optimized for each ATP concentration. Compound IC₅₀ values were determined by 12-point inhibition curves (from 50 to 0.000282 μ M) in duplicate. **Structure determination.** Before crystallization, 0.1 mM of EGFR(T790M/V948R) was incubated for 1 h with 0.5 mM EAI001, 1 mM adenosine 5'-(β , γ -imido) triphosphate (AMP-PNP) and 10 mM MgCl₂ at room temperature. Crystals of EGFR(T790M/V948R) in complex with EAI001 were prepared by hanging-drop vapour diffusion method over a reservoir solution containing 0.1 M Bis-Tris (pH 5.5), 25% PEG 3350, 5 mM tris (2-carboxyethyl)-phosphine (TCEP). Crystals were flash-frozen in liquid nitrogen after rapid immersion in a cryoprotectant solution containing 0.1 M Bis-Tris 5.5, 25% PEG3350, 10% ethylene glycol and 5 mM TCEP. Diffraction data were recorded using a Mar343 image plate detector on a rotating anode source at 100 K. Data were processed and merged as described previously¹⁴. The structure was determined by molecular replacement with the program PHASER using an inactive EGFR kinase structure (PDB, 2GS7) as the search model. Repeated rounds of manual refitting and crystallographic refinement were performed using COOT and REFMAC. The inhibitor was modelled into the closely fitting positive $F_o - F_c$ electron density and then included in following refinement cycles. Although the EAI001 preparation used in crystallization was racemic, the density clearly corresponded to the *R* stereoisomer and was modelled accordingly. Topology and parameter files for the inhibitors were generated using PRODRG. Statistics for diffraction data processing and structure refinement are shown in Extended Data Table 3.

Tissue Culture. Cells were maintained in 10% FBS/RPMI supplemented with 100 μ g ml⁻¹ penicillin/streptomycin (Hyclone SH30236.01). The cells were collected with 0.25% trypsin/EDTA (Hyclone SH30042.1), re-suspended in 5% FBS/RPMI penicillin/streptomycin and plated at 7,500 cells per well in 50 μ l of media in a 384-well black plate with clear bottoms (Greiner 789068G). The cells were allowed to incubate overnight in a 37 °C, 5% CO₂ humidified tissue culture incubator. The 12-point serial diluted test compounds were transferred to the plate containing cells by using a 50 nl Pin Head device (Perkin Elmer) and the cells were placed back in the incubator for 3 h. All cell lines were tested and found negative for mycoplasma contamination using the MycoAlert Mycoplasma Detection Kit (Lonza).

Phospho-EGFR (Y1173) target modulation assay. HaCaT cells were stimulated with 10 ng ml⁻¹ EGF (Peprotech AF-100-15) for 5 min at room temperature. Constitutively activated EGFR mutant cell lines (H1975 and H3255) were not stimulated with EGF. The media was reduced to 20 μ l using a Bio-Tek ELx405 Select plate washer. Cells were lysed with 20 μ l of 2 \times lysis buffer containing protease and phosphatase inhibitors (2% Triton X-100, 40 mM Tris (pH 7.5), 2 mM EDTA, 2 mM EGTA, 300 mM NaCl, 2 \times complete cocktail inhibitor (Roche 11 697 498 001), 2 \times phosphatase inhibitor cocktail set II and set III (Sigma P5726 and P0044)). The plates were shaken for 20 min. An aliquot of 25 μ l from each well was transferred to prepared ELISA plates for analysis.

For the experiment studying the effect of EGF pre-treatment on EAI045 target modulation, H1975 cells were collected and plated in 0.5% FBS/RPMI penicillin/streptomycin. On the following day, cells were pre-treated with 0.5% FBS/RPMI media with or without 10 ng EGF per ml for 5 min. Compound was added and assay was carried out as described above. The experiment was performed twice with duplicate samples in each experiment.

Phospho-EGFR (Y1173) ELISA. Solid white 384-well high-binding ELISA plates (Greiner 781074) were coated with 5 μ g ml⁻¹ goat anti-EGFR capture antibody overnight in 50 mM carbonate/bicarbonate (pH 9.5) buffer. Plates were blocked with 1% BSA (Sigma A7030) in PBS for 1 h at room temperature, and washes were carried out with a Bio-Tek ELx405 Select using four cycles of 100 μ l TBS-Tween

(20 mM Tris, 137 mM NaCl, 0.05% Tween-20) per well. A 25 μ l aliquot of lysed cell was added to each well of the ELISA plate and incubated overnight at 4 °C with gentle shaking. After washing, 1:1,000 anti-phospho-EGFR in 0.2% BSA/TBS-Tween was added and incubated for 2 h at room temperature. After washing, 1:2,000 anti-rabbit-HRP (horseradish peroxidase) in 0.2% BSA/TBS-Tween was added and incubated for 1 h at room temperature. Chemiluminescent detection was carried out with SuperSignal ELISA Pico substrate. Luminescence was read with an EnVision plate reader.

Western blotting. Cell lysates were equalized to protein content determined by Coomassie Plus protein assay reagent (ThermoScientific 1856210) and loaded onto 4–12% NuPAGE Bis-Tris gels with MOPS running buffer with LDS Sample buffer supplemented with DTT. Gel proteins were transferred to PVDF membranes with an iBlot Gel Transfer Device. 1 \times Casein-blocked membranes were probed with primary antibodies overnight at 4 °C on an end-over-end rotisserie. Membranes were washed with TBS-Tween and HRP-conjugated secondary antibodies were added for 1 h at room temperature. After washing, HRP was detected using Luminata Forte Western HRP Substrate reagent and recorded with a Bio-Rad VersaDoc imager.

H1975, H3255 and HaCaT proliferation assays. H1975, H3255 and HaCaT cell lines were plated in solid white 384-well plates (Greiner) at 500 cells per well in 10% FBS RPMI penicillin/streptomycin media. Using a Pin Tool, 50 nl of serial diluted compounds were transferred to the cells. After 3 days, cell viability was measured by CellTiter-Glo (Promega) according to manufacturer's instructions. Luminescent readout was normalized to 0.1% DMSO-treated cells and empty wells. Data was analysed by nonlinear regression curve fitting and EC₅₀ values were reported.

Ba/F3 cell proliferation models. The EGFR mutant L858R, L858R/T790M, delE746_A750/T790M, L858R/T790M/C797S and del/T790M/C797S Ba/F3 cells have been previously described¹⁵. The EGFR(I941R) mutation was introduced via site directed mutagenesis using the Quick Change Site-Directed Mutagenesis kit (Stratagene) according to the manufacturer's instructions. All constructs were confirmed by DNA sequencing. The constructs were shuttled into the retroviral vector JP1540 using the BD Creator System (BD Biosciences). Ba/F3 cells were infected with retrovirus and according to standard protocols, as described previously³⁰. Stable clones were obtained by selection in puromycin (2 μ g ml⁻¹). Ba/F3 cells have not been authenticated as there is no publicly available fingerprint for Ba/F3 cells. All variants used were confirmed to contain the correct EGFR mutation by sequencing. All Ba/F3 cells were tested for mycoplasma contamination and confirmed to be free of contamination.

Growth and inhibition of growth was assessed by MTS assay and was performed according to previously established methods¹⁵. Ba/F3 cells of different EGFR genotypes were exposed to treatment for 72 h and the number of cells used per experiment determined empirically and has been previously established¹⁵. All experimental points were set up in six wells and all experiments were repeated at least three times. The data was graphically displayed using GraphPad Prism version 5.0 for Windows, (GraphPad software; <http://www.graphpad.com>). The curves were fitted using a nonlinear regression model with a sigmoidal dose response.

NIH-3T3 cell studies. NIH-3T3 cells were infected with retroviral constructs expressing EGFR mutants according to standard protocols, as described previously^{15,19}. Stable clones were obtained by selection in puromycin (2 μ g ml⁻¹).

Mouse efficacy studies. EGFR(TL) (bearing L858R/T790M point mutations) and EGFR(TD) (bearing exon19del/T790M point mutations) mice were generated as previously described^{15,27}. The EGFR(L858R/T790M/C797S) (denoted as TLCS hereafter) mutant mouse cohort was established briefly as follows: the full-length human TLCS cDNA was generated by site-directed mutagenesis using the Quickchange site directed mutagenesis kit (Agilent Technologies) and further verified by DNA sequencing. Sequence-verified targeting vectors were co-electroporated with an FLPe recombinase plasmid into v6.5 C57BL/6J (female) \times 129/sv (male) embryonic stem cells (Open Biosystems) as described elsewhere³¹. Resulting hygromycin-resistant embryonic stem clones were evaluated for transgene integration via PCR. Then, transgene-positive embryonic stem clones were injected into C57BL/6 blastocysts, and the resulting chimaeras were mated with BALB/c wild type mice to determine germline transmission of the TLCS transgene. Further detail on the generation and characterization of the TLCS transgenic mice is provided in Supplementary Fig. 3. Progeny of TL, TD and TLCS mice were genotyped by PCR of tail DNA. The TL and TD mice were fed a doxycycline diet at 6 weeks of age to induce EGFR(TL) or EGFR(TD) expression, respectively. The TLCS mice were intranasally instilled with Ad-Cre (University of Iowa viral vector core) at 6 weeks of age to excise the loxP sites, activating EGFR(TLCS) expression.

The EAI045 compound was dissolved in 10% NMP (10% 1-methyl-2-pyrrolidinone: 90% PEG-300), and was dosed at 60 mg kg⁻¹ daily by oral gavage. Cetuximab was administered at 1 mg mouse⁻¹ every other day by intraperitoneal injection. The TL, TD and TLCS mice were monitored by MRI to quantify lung tumour

burden before being assigned to various study treatment cohorts, which were non-blinded and not formally randomized. All treated mice had an equal initial tumour burden. MRI evaluation was repeated every 2 weeks during treatment. The animals were imaged with a rapid acquisition with relaxation enhancement sequence (repetition time = 2000 ms; echo time = 25 ms) in the coronal and axial planes with a 1-mm slice thickness and with respiratory gating. The detailed procedure for MRI scanning has been previously described²⁷. The tumour burden volumes were quantified using 3-dimensional Slicer software. Source data for tumour volume measurements are provided in Supplementary Fig. 2.

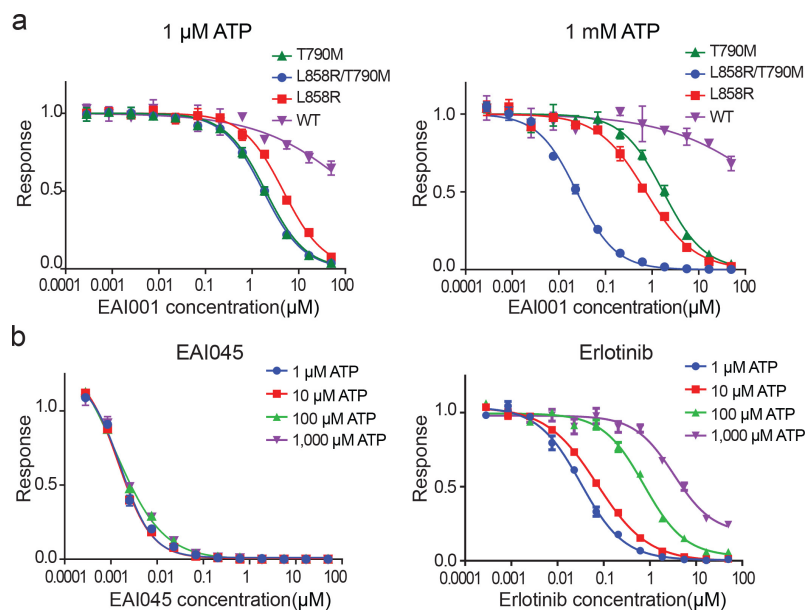
All care of experimental animals was in accordance with Harvard Medical School/Dana-Farber Cancer Institute institutional animal care and use committee (IACUC) guidelines. All mice were housed in a pathogen-free environment at a DFCI animal facility and handled in strict accordance with Good Animal Practice as defined by the Office of Laboratory Animal Welfare. None of the tumour efficacy experiments presented in this manuscript exceeded the 2 cm maximal diameter tumour size, as permitted by the Dana-Farber Cancer Institute IACUC.

Synthesis and characterization of EAI045. 2-(5-fluoro-2-hydroxyphenyl)-2-(1-oxo-2,3-dihydro-1H-isindol-2-yl)-N-(1,3-thiazol-2-yl)acetamide (EAI045)

was prepared from 2-amino-2-(5-fluoro-2-methoxyphenyl)acetic acid using a reaction sequence similar to that previously described³² followed by demethylation with boron tribromide.

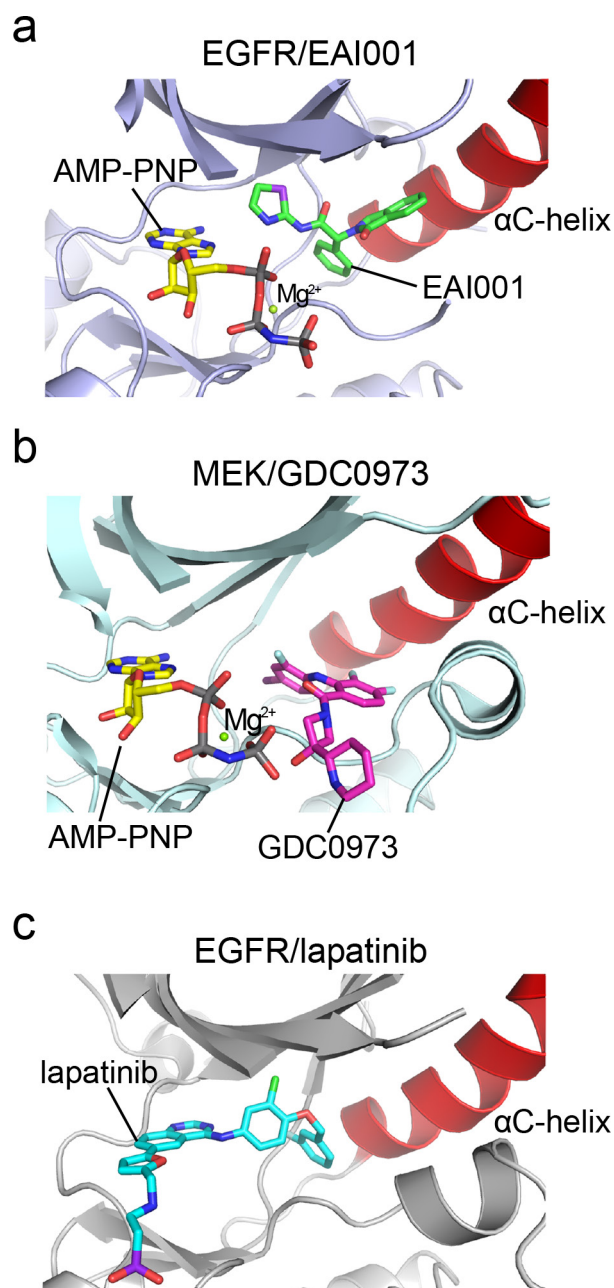
¹H NMR (400 MHz, DMSO-d₆) δ 12.61 (s, 1H), 9.96 (s, 1H), 7.73 (d, J = 7.5 Hz, 1H), 7.66–7.54 (m, 2H), 7.52 (dd, J = 1.0, 7.4 Hz, 1H), 7.49 (d, J = 3.6 Hz, 1H), 7.27 (d, J = 3.5 Hz, 1H), 7.11 (td, J = 3.2, 8.6 Hz, 1H), 6.90 (dd, J = 4.8, 8.9 Hz, 1H), 6.85 (dd, J = 3.1, 9.2 Hz, 1H), 6.31 (s, 1H), 4.61 (d, J = 17.5 Hz, 1H), 3.98 (d, J = 17.5 Hz, 1H); ¹⁹F NMR (376 MHz, DMSO-d₆) δ –125.15 (s, 1F); LCMS: Rt 1.278 min; ESMS m/z 384.20 (M⁺H⁺).

29. Hong, L., Quinn, C. M. & Jia, Y. Evaluating the utility of the HTRF Transcreeper ADP assay technology: a comparison with the standard HTRF assay technology. *Anal. Biochem.* **391**, 31–38 (2009).
30. Engelman, J. A. *et al.* ErbB-3 mediates phosphoinositide 3-kinase activity in gefitinib-sensitive non-small cell lung cancer cell lines. *Proc. Natl Acad. Sci. USA* **102**, 3788–3793 (2005).
31. Beard, C., Hochedlinger, K., Plath, K., Wutz, A. & Jaenisch, R. Efficient method to generate single-copy transgenic mice by site-specific integration in embryonic stem cells. *Genesis* **44**, 23–28 (2006).
32. Muller, G. W. Cyclic amides. *US patent* 5698579 (1997).

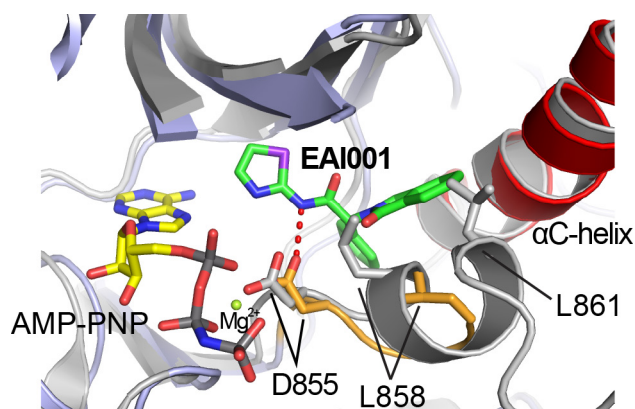


Extended Data Figure 1 | Inhibition of wild-type and mutant EGFR kinases by EAI001 and EAI045 in purified enzyme assays. a, Inhibition of wild-type and mutant EGFR kinases by EAI001. Activity of the indicated mutant EGFR kinase (residues 696–1022) was measured in the presence of increasing concentrations of EAI001. The HTRF assay was

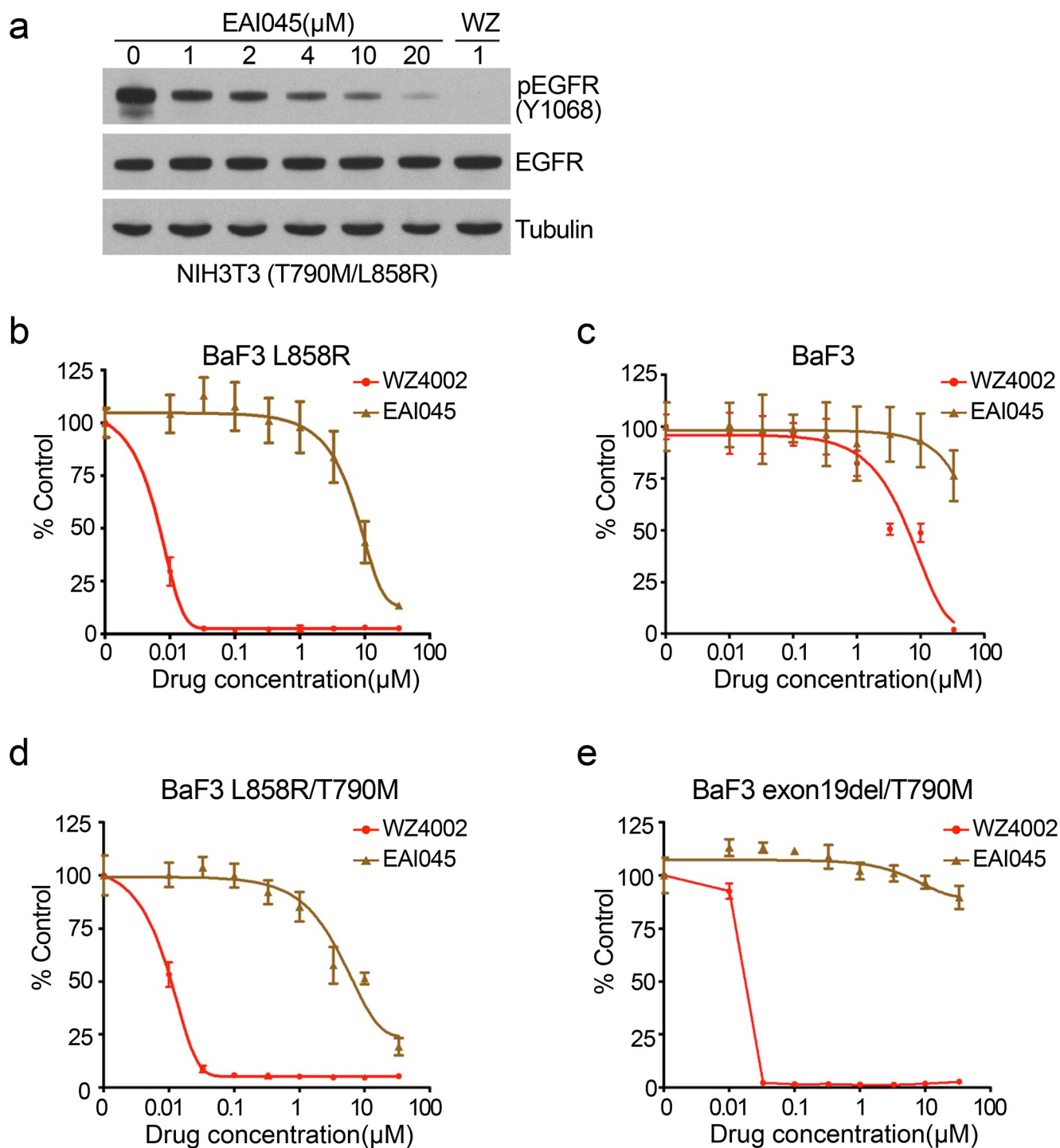
carried out using either 1 μM ATP (left) or 1 mM ATP (right). **b,** Inhibition of EGFR(L858R/T790M) by EAI045 (left) or erlotinib (right) at a range of ATP concentrations, as indicated. Assay was performed using an HTRF-based assay as described in the Methods. Error bars indicate s.d. ($n = 2$).



Extended Data Figure 2 | Comparison of the binding site of EGFR allosteric inhibitors with those of lapatinib and allosteric MEK inhibitors. **a**, Structure of EAI001 in complex with EGFR for comparison. **b**, Structure of MEK1 kinase bound to allosteric inhibitor GDC0973 (PDB, 4AN2). GDC0973 (also called XL518, cobimetinib) and other allosteric MEK inhibitors occupy a pocket created by displacement of the C-helix in the inactive conformation of the kinase. Most allosteric MEK inhibitors make hydrogen-bond interactions with the γ -phosphate group of ATP that are important for their potency. The allosteric EGFR inhibitors we describe here bind in a generally analogous location in EGFR, but lack any clear structural similarity to MEK inhibitors and do not contact the γ -phosphate group of ATP. **c**, The structure of lapatinib bound to EGFR (PDB, 1XKK). Both lapatinib and neratinib (see Fig. 1d) bind an inactive conformation of the kinase. Like gefitinib and erlotinib, both occupy the ATP site, but also extend into the allosteric pocket occupied by EAI001. Note that like neratinib, lapatinib places aromatic phenyl or pyridinyl groups in positions similar to those occupied by the aminothiazole and phenyl substituents of EAI001.

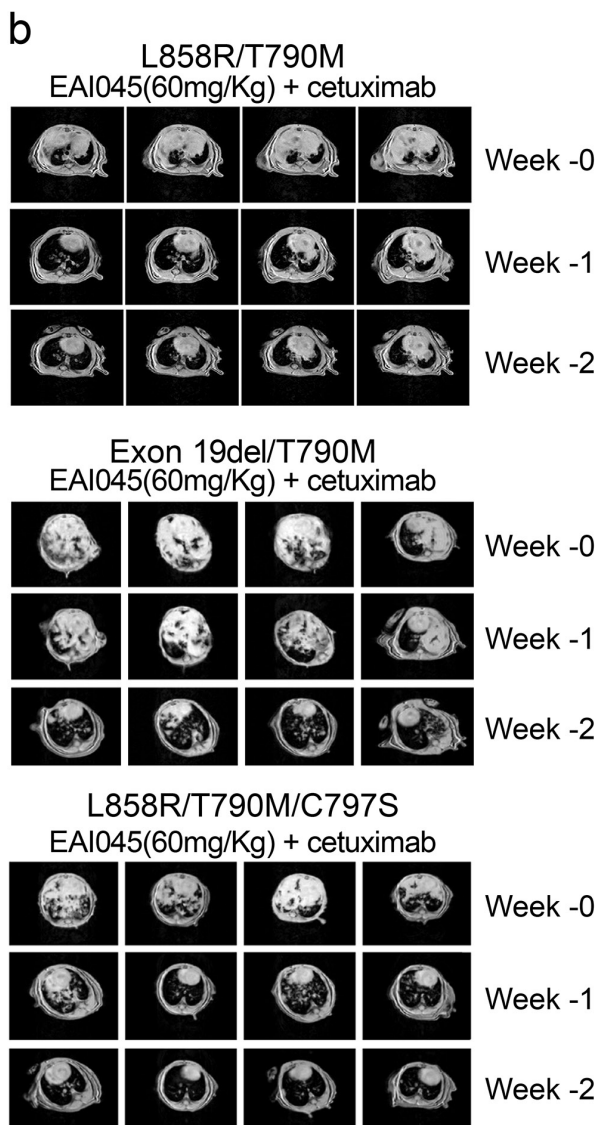
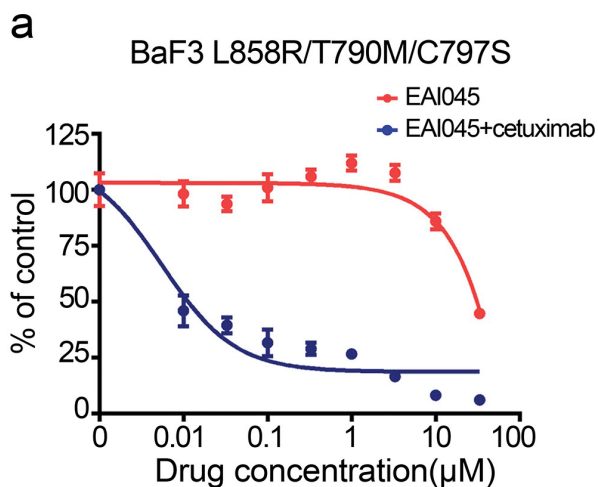


Extended Data Figure 3 | EAI001 binding is incompatible with the inactive conformation of wild-type EGFR. Superposition of the EAI001-bound EGFR structure reported here with the structure of wild-type EGFR kinase in the inactive conformation (grey, PDB, 2GS7). EAI001 (shown with carbon atoms in green) clashes with the side chains of leucines 858 and 861 in the wild-type EGFR structure. These leucine residues lie in a short helical segment at the N terminus of the activation loop. The L858R substitution disrupts this helix. We propose that this effect explains, in part, the selectivity of the allosteric inhibitor for the L858R/T790M mutant. Note that EAI001 was crystallized with the EGFR(T790M/V948R), as we were unable to obtain crystals with the L858R/T790M or L858R/T790M/V948R proteins. The compound induces unstructuring of the activation loop helix and repositions L858, which is in contact with the 1-oxoisindolinyl group of the inhibitor. The location and conformation of the inhibitor is expected to be the same in the context of the L858R mutation, but the details of the interaction with this portion of the activation loop will necessarily differ due to the mutation.



Extended Data Figure 4 | Cellular activity of EAI045. a, EAI045 inhibition of EGFR(L858R/T790M) in NIH-3T3 cells. Western blotting with the indicated concentrations of the allosteric inhibitor or with 1 μ M WZ4002 as control (WZ) was carried out 6 h after compound addition.

b–e, Profiling of EAI045 in Ba/F3 models bearing mutant EGFR or the parental Ba/F3 cell line, as indicated. Inhibition by WZ4002 is shown as a positive control. For gel source data, see Supplementary Fig. 1.



Extended Data Figure 5 | Cellular and *in vivo* efficacy of EAI045 in combination with cetuximab. **a**, Ba/F3 cells bearing EGFR(L858R/T790M/C797S) were treated with EAI045 alone or with EAI045 plus cetuximab and proliferation was measured using the MTS assay after 72 h. **b**, MRI imaging of cohorts L858R/T790M, exon19del/T790M, and L858R/T790M/C797S genetically engineered EGFR-mutant mice before treatment and 1 or 2 weeks after treatment with EAI045 and cetuximab. These cohorts of tumour bearing mice were used for short term efficacy and pharmacodynamic studies, and are distinct from those used for the tumour volume measurements shown in Fig. 3.

Extended Data Table 1 | The selectivity of EAI045 on a panel of kinases

Kinase	% inhibition	Kinase	% inhibition	Kinase	% inhibition	Kinase	% inhibition	Kinase	% inhibition	Kinase	% inhibition
ABL1	1	CK1-EPSILON	0	FMS	0	MAP4K5	13	PAK4	0	ROCK1	0
AKT1	0	CK1-GAMMA1	2	FRAP1	0	MAPK1	0	PAK5	0	ROCK2	0
AKT2	0	CK1-GAMMA2	0	FYN	0	MAPK3	0	PAK6	0	RON	0
AKT3	0	CK1-GAMMA3	6	GRK3	0	MAPKAPK2	0	PASK	1	ROS	1
ALK2	0	CK2	0	GRK5	0	MAPKAPK3	0	PDGFR-ALPHA	0	RSK1	0
ALK5	0	CLK1	0	GRK6	0	MARK1	0	PDGFR-BETA	4	RSK2	0
ALK6	0	CLK2	1	GRK7	0	MARK3	0	PDK1	0	RSK3	0
AMP-A1B1G1	1	CLK3	0	GSK-3-ALPHA	6	MARK4	0	PHK-GAMMA1	0	RSK4	0
AMP-A2B1G1	0	CLK4	0	GSK-3-BETA	0	MEK1	0	PHK-GAMMA2	0	SGK1	6
ARG	2	CRAF	0	HASPIN	1	MEK2	0	PI3K-ALPHA	0	SGK2	0
ARK5	0	CSK	0	HCK	0	MELK	0	PI3K-BETA	0	SGK3	0
AURORA-A	0	DAPK1	0	HIPK1	1	MER	8	PI3K-DELTA	0	SIK	0
AURORA-B	4	DAPK3	1	HIPK2	2	MET	0	PI3K-GAMMA	0	SLK	0
AURORA-C	0	DCAMKL2	0	HIPK3	0	MKNK1	0	PI4-K-BETA	9	SNF1LK2	0
AXL	4	DDR2	4	HIPK4	0	MNK2	3	PIM1	0	SPHK1	0
BLK	3	DYRK1A	0	IGF1R	4	MRCK-ALPHA	0	PIM2	2	SPHK2	0
BMX	1	DYRK1B	2	IKK-ALPHA	16	MRCK-BETA	0	PIM3	1	SRC	0
BRAF	0	DYRK3	2	IKK-BETA	0	MSK1	0	PKA	3	SRMS	0
BRK	0	DYRK4	1	IKK-EPSILON	0	MSK2	2	PKACB	0	SRPK1	0
BRSK1	0	EGFR	7	INSR	0	MSSK1	0	PKC-ALPHA	0	SRPK2	10
BRSK2	0	EPH-A1	0	IRAK1	4	MST1	0	PKC-EPSILON	14	STK16	0
BTk	0	EPH-A2	7	IRAK4	0	MST2	0	PKC-ETA	0	SYK	0
CAMK1A	0	EPH-A3	1	IRR	0	MST3	0	PKC-GAMMA	0	TAK1-TAB1	0
CAMK1D	0	EPH-A4	0	ITK	0	MST4	0	PKC-IOTA	0	TAOK2	0
CAMK2A	2	EPH-A5	4	JAK1	0	MUSK	0	PKC-THETA	0	TAOK3	0
CAMK2B	0	EPH-A8	0	JAK2	2	NDR1	2	PKC-ZETA	0	TBK1	3
CAMK2D	4	EPH-B1	0	JAK3	0	NDR2	0	PKN1	0	TEC	4
CAMK2G	9	EPH-B2	0	JNK1	0	NEK1	0	PKN2	1	TIE2	0
CAMK4	1	EPH-B3	0	JNK2	0	NEK2	2	PLK1	7	TNIK	0
CDK1	2	EPH-B4	0	JNK3	0	NEK3	0	PLK3	5	TNK1	0
CDK2-CYCLINA	1	ERB-B2	4	KDR	0	NEK6	0	PLK4	0	TNK2	2
CDK2-CYCLINE	0	ERB-B4	16	KIT	0	NEK7	0	PRAK	0	TRKA	0
CDK3-CYCLINE	0	FER	4	LATS2	2	NEK9	0	PRKD1	0	TRKB	3
CDK4-CYCLIND	0	FES	2	LCK	0	P38-ALPHA	0	PRKD2	2	TRKC	4
CDK5	0	FGFR1	5	LIMK1	0	P38-BETA	0	PRKD3	0	TSSK1	2
CDK5-P25	0	FGFR2	1	LOK	1	P38-DELTA	0	PRKG1	0	TSSK2	0
CDK6-CYCLIND3	2	FGFR3	1	LRRK2-G2019S	0	P38-GAMMA	2	PRKG2	8	TTK	0
CDK7	0	FGFR4	0	LTK	1	P70S6K1	0	PRKX	3	TXK	0
CDK9-CYCLINT1	6	FGR	0	LYNA	9	P70S6K2	0	PTK5	0	TYK2	0
CHEK1	0	FLT-1	8	LYNB	3	PAK1	0	PYK2	4	TYRO3	0
CHEK2	0	FLT-3	0	MAP4K2	0	PAK2	0	RET	0	YES	0
CK1	0	FLT-4	1	MAP4K4	2	PAK3	0	RIPK2	0	ZAP70	1

*Percent inhibition was measured in the presence of 1 μ M EAI045. Experiment was performed once with duplicate samples.

Extended Data Table 2 | Selectivity of EAI045 against a panel of non-kinase targets

Assay Name	IC ₅₀ (μM)
Adenosine 2A receptor binding assay	>30
Adenosine 3 receptor binding assay	>30
Adrenergic Alpha 2C receptor assay	>30
Alpha1A adrenergic calcium flux assay (agonist mode)	>30
Alpha1A adrenergic calcium flux assay (antagonist mode)	>30
Beta 1 adrenergic receptor assay	>30
COX-1 assay	>30
CYP3A4 Induction Reporter Gene	>10
Dopamine D2 receptor assay	>30
Dopamine Transporter assay	>30
H1 receptor calcium assay (agonist mode)	>30
H1 receptor calcium assay (antagonist mode)	>30
Histamine H1 receptor assay	>30
Melanocortin MC3 receptor binding assay	>30
Monoamine Oxidase A assay	>30
Muscarinic M1 receptor assay	>30
Muscarinic M2 calcium flux assay with ATP priming (agonist mode)	>30
Nicotinic (CNS) Receptor binding (human IMR32 cells)	>30
Norepinephrine Transporter assay	>30
PPARgamma Receptor agonist assay	>30
PPARgamma Receptor antagonist assay	>30
PXR Receptor agonist assay	16
PXR Receptor antagonist assay	>3
Phosphodiesterase 3 assay (human platelets)	>30
Phosphodiesterase 4D assay	>30
Pregnane X Receptor (PXR; SXR) binding assay	7
Progesterone Receptor agonist assay	>30
Progesterone Receptor antagonist assay	>30
Serotonin 5HT2A calcium flux assay (agonist mode)	>30
Serotonin 5HT2A calcium flux assay (antagonist mode)	>30

Extended Data Table 3 | Crystallographic data collection and refinement statistics

Crystal name		EGFR T790M/V948R EAI001
Data collection		
Space group		C2
Cell dimensions		
a, b, c (Å)		155.1, 72.5, 76.0
α , β , γ (°)		90, 113.2, 90
Resolution (Å)		42.24 - 2.31 (2.39)*
R _{merge}		0.11 (0.51)
I/ σ		10.4 (1.9)
Completeness (%)		97.8 (94.0)
Redundancy		3.4 (2.9)
Refinement		
Resolution (Å)		42.24 - 2.31
No. Reflections		33377
R _{work} / R _{free}		0.174/0.206
No. Atoms		
Protein		4826
Ligand/ion (AMPPNP/EAI001/Mg ²⁺)		62/25/2
Water		398
B-factors		
Protein		33.70
Ligand/ion (AMPPNP/EAI001/Mg ²⁺)		25.90
Water		36.60
R.m.s deviations		
Bond lengths (Å)		0.008
Bond angles (°)		1.136

Diffraction data were recorded from a single crystal.

*Values in parentheses are for highest resolution shell.

Extended Data Table 4 | Cellular activity of EAI045 in lung cancer cell lines

Cell line (EGFR)	EAI045 EC ₅₀ (μM)	
	Target modulation*	Proliferation†
H1975 (L858R/T790M)	0.002 (4)	>10 (2)
H3255 (L858R)	0.163 (4)	>10 (2)
HaCaT (WT)	>10 (2)	>10 (1)

*ELISA-based assay for phosphorylation of EGFR Y1173; the number of times each experiment was repeated is given in parentheses.

†The number of times each experiment was repeated is given in parentheses.

Diverse roles of assembly factors revealed by structures of late nuclear pre-60S ribosomes

Shan Wu¹, Beril Tutuncuoglu², Kaige Yan¹, Hailey Brown², Yixiao Zhang¹, Dan Tan^{3,4}, Michael Gamalinda², Yi Yuan¹, Zhifei Li¹, Jelena Jakovljevic², Chengying Ma¹, Jianlin Lei¹, Meng-Qiu Dong^{3,4}, John L. Woolford Jr² & Ning Gao¹

Ribosome biogenesis is a highly complex process in eukaryotes, involving temporally and spatially regulated ribosomal protein (r-protein) binding and ribosomal RNA remodelling events in the nucleolus, nucleoplasm and cytoplasm^{1,2}. Hundreds of assembly factors, organized into sequential functional groups^{3,4}, facilitate and guide the maturation process into productive assembly branches in and across different cellular compartments. However, the precise mechanisms by which these assembly factors function are largely unknown. Here we use cryo-electron microscopy to characterize the structures of yeast nucleoplasmic pre-60S particles affinity-purified using the epitope-tagged assembly factor Nog2. Our data pinpoint the locations and determine the structures of over 20 assembly factors, which are enriched in two areas: an arc region extending from the central protuberance to the polypeptide tunnel exit, and the domain including the internal transcribed spacer 2 (ITS2) that separates 5.8S and 25S ribosomal RNAs. In particular, two regulatory GTPases, Nog2 and Nog1, act as hub proteins to interact with multiple, distant assembly factors and functional ribosomal RNA elements, manifesting their critical roles in structural remodelling checkpoints and nuclear export. Moreover, our snapshots of compositionally and structurally different pre-60S intermediates provide essential mechanistic details for three major remodelling events before nuclear export: rotation of the 5S ribonucleoprotein, construction of the active centre and ITS2 removal. The rich structural information in our structures provides a framework to dissect molecular roles of diverse assembly factors in eukaryotic ribosome assembly.

Assembly of pre-60S ribosomes occurs in consecutive stages, orchestrated by coordinated groups of assembly factors. The presence or absence of three mostly non-overlapping factors in pre-60S particles, Nsa1, Nog2 and Nmd3, defines a continuous transition from the nucleolus through the nucleoplasm to final stages licensing nuclear export (Extended Data Fig. 1). Nog2, an essential GTPase⁵, enters pre-60S particles in the nucleolus, and is present during most nucleoplasmic stages. The lifetime of Nog2 coincides with three important pre-60S remodelling and processing events: rotation of the 5S ribonucleoprotein (RNP)⁶, construction of the active site and cleavage of ITS2 (ref. 5), as well as the temporally regulated binding and release of assembly factors⁷. Nog2 departure constitutes a critical checkpoint for nuclear export of pre-60S particles⁸. The remodelling ATPase Rea1 is thought to catalyse conformational changes in late nucleoplasmic particles that stimulate the GTPase activity of Nog2 (ref. 8). This enables release of Nog2 and replacement by the key export factor Nmd3, whose binding site overlaps with that of Nog2 (refs 8, 9). This model of nucleoplasmic pre-60S maturation was established largely from biochemical and genetic experiments (reviewed in ref. 1). Low-resolution cryo-electron microscopy (cryo-EM) maps have revealed the location of several different assembly factors^{6,10–14}, but atomic contacts with the 60S

subunit are only known for Tif6 (ref. 15), Arx1, Alb1 and Reil (ref. 16). Nevertheless, spatial relationships among most of the assembly factors in pre-60S particles remain unclear. In particular, key assembly events responsible for activating successive maturation checkpoints are yet to be determined.

To further explore the mechanism of late nuclear steps in pre-60S assembly, we characterized structures of native nucleoplasmic particles isolated from *Saccharomyces cerevisiae*, using epitope-tagged assembly factor Nog2 (Extended Data Fig. 1a). These Nog2-particles were subjected to cryo-electron microscopy (cryo-EM) (Extended Data Table 1) to determine a series of structures (hereafter termed states 1–3) (Extended Data Fig. 2), presumably reflecting temporally related snapshots of final maturation steps of pre-60S particles before nuclear export. One of these structures, state 1, was solved at a nominal resolution of 3.08 Å (Extended Data Fig. 3b, c and Supplementary Video 1). We identified over 30 assembly factors in Nog2-particles (Extended Data Fig. 1b). Guided by chemical cross-linking of proteins coupled with mass spectrometry (XL-MS) (Supplementary Table 1), we were able to build atomic models for 19 of these assembly factors in the density map of state 1 (Fig. 1 and Extended Data Fig. 4). Intriguingly, 14 of them are located in the arc region of the central protuberance–polypeptide tunnel exit on the intersubunit surface (Fig. 1a), and five are immediately adjacent or bound to ITS2 (Fig. 1a). In addition, we also located Sda1, Rea1 and the Rix1 sub-complex¹⁴ in the map of state 2.

Nog2 binds at the centre of the pre-60S particle, via interaction of its GTPase domain and carboxy (C)-terminal domain with a multi-helical junction (Fig. 2), making extensive contacts with H93, H62, H64, H67, H69 and H71 of 25S ribosomal RNA (rRNA), and Bud20 (Extended Data Fig. 5g). This interaction stabilizes H69 and H71 in a nearly 180°-flipped position (Fig. 2c–e) compared with their mature forms¹⁷. In addition, the C-terminal extended loop of Rpf2 (residues 275–300) is inserted into the interface of Nog2–GTPase domain–C-terminal domain and H69–H71 (Extended Data Fig. 5a), also contributing to the displacement of H69–H71. Unexpectedly, the amino (N)-terminal extension of Nog2 (residues 1–200) lacks tertiary structures (Fig. 2a). Instead, in its fully extended form, the N-terminal extension of Nog2 wanders around the inner surface of the tRNA passageway on the pre-60S particle and interacts with multiple components, including H92, L23, Nog1, H43, Rsa4, Nsa2, Rpf2, Rrs1 and H86 (Extended Data Fig. 5). The very N terminus of Nog2 ends at a helical junction composed of H68, H74, H75 and H93 (Extended Data Fig. 5h). These observations nicely explain the previous model that Nog2 represents a converging node for different assembly branches and that its recruitment requires the prior association of multiple assembly factors^{18,19}.

Our structures also reveal potential functions for the GTPase Nog1, which docks to a similar position in the pre-60S particle as its homologue ObgE does in bacterial large ribosomal subunits²⁰, with

¹Ministry of Education Key Laboratory of Protein Sciences, Beijing Advanced Innovation Center for Structural Biology, School of Life Sciences, Tsinghua University, Beijing 100084, China.

²Department of Biological Sciences, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213, USA. ³Graduate Program in Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing 100730, China. ⁴National Institute of Biological Sciences, Beijing 102206, China.

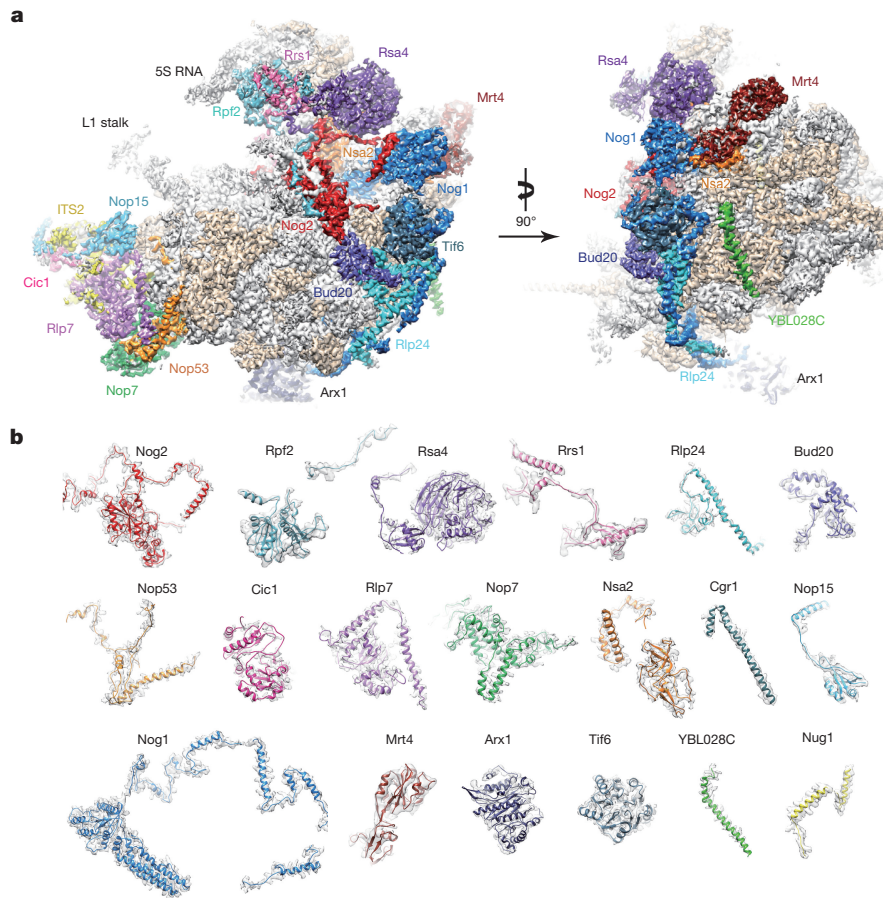


Figure 1 | Cryo-EM structure (state 1) of the pre-60S particle purified from epitope-tagged Nog2. a, The 3.08-Å cryo-EM map of state 1 is displayed in surface representation, with density of each assembly factor separately coloured. The 25S rRNA and r-proteins are coloured grey and

its N-terminal four-helical-bundle domain (NTD) pointing to the peptidyl transferase centre (Fig. 3a). Interestingly, the NTD of Nog1 directly passes through H89, separating it into two strands (Fig. 3a–c). Besides rRNA, the GTPase domain and NTD of Nog1 also interact with Nog2 and Nsa2, respectively (Extended Data Fig. 6a). The C-terminal extension (CTE) of Nog1, similar to the N-terminal extension of Nog2, wraps around the pre-60S particle by over one-quarter of its circumference. On its way from the P0 stalk base to the polypeptide tunnel exit, the CTE of Nog1 makes extensive contacts with nearly all of the assembly factors and r-proteins in this arc region (Tif6, Rlp24, Arx1, L3, L31, L22, L19, L35) (Fig. 3d) and with a variety of rRNA helices. The spatial relationship of Nog1 with these assembly factors agrees well with the previous model for ordered recruitment and release of assembly factors during cytoplasmic maturation of pre-60S particles¹. In particular, the CTE of Nog1 interlocks with Rlp24 by wrapping around a long helix (residues 85–130) at the C-terminal end of Rlp24 (Fig. 3d and Extended Data Fig. 6b), suggesting that these two assembly factors might be recruited and released as a subcomplex⁷. Indeed, the release of Nog1 and replacement of Rlp24 with L24 in the cytoplasm is catalysed by the ATPase Drg1 (refs 21, 22). Surprisingly, at the polypeptide tunnel exit, the CTE of Nog1 turns into the polypeptide tunnel, extending all the way through the tunnel to the peptidyl transferase centre (Fig. 3e). It is tempting to hypothesize that this C terminus of Nog1 (~75 residues) might enable polypeptide exit tunnel assembly, and/or test-drive the tunnel by surveying the conformational status of tunnel wall components (such as L39, L17 and L4). Altogether, our data suggest several distinct roles for Nog1 in the maturation of pre-60S particles. While the NTD might serve to remodel the peptidyl transferase centre, the CTE of Nog1 apparently

acts as a scaffold for assembly of many assembly factors and r-proteins, and might participate in quality control of polypeptide tunnel construction. Notably, a recent study showed that the tunnel is again probed in the cytoplasm by Rei1 (ref. 16) in a similar fashion as the CTE of Nog1 does (Extended Data Fig. 6c, d), indicating the existence of continuous proofreading of the ribosomal tunnel from the nucleolus to the cytoplasm.

In the map of state 1, a portion of the ITS2 pre-rRNA spacer is well resolved. This includes 59 nucleotides extending from the 3'-end of 5.8S rRNA, and 6 nucleotides of ITS2 at the 5'-end of 25S rRNA (Fig. 4a–c), consistent with the presence of 25.5S and 7S pre-rRNAs in Nog2-particles⁵. In addition to known ITS2-binding factors Nop15, Rlp7 and Cic1 (ref. 23), we also identified Nop7 and Nop53 (ref. 24) in the region of ITS2 (Fig. 4d). This close co-localization of Nop15, Rlp7, Cic1, and Nop7 around ITS2 explains their mutually interdependent association with pre-60S particles¹. Notably, Nop53 is required to recruit Mtr4 which participates in exosome-mediated ITS2 removal^{18,25}. Three r-proteins, L8, L25 and L27, also interact with these ITS2 factors (Extended Data Fig. 7). L8 directly contacts Nop15, Cic1 and Nop7 (Extended Data Fig. 7a–d), complementing previous data that L8 is required for the assembly of these A3 factors²⁶. The high protein content in the ITS2 region further suggests that these factors function to chaperone and protect ITS2 for proper processing. Given the space required for progressive trimming of 7S pre-rRNA from its 3'-end by the exosome and other nucleases^{27,28}, it is conceivable that de-coating of assembly factors from ITS2 is coordinated with stepwise removal of ITS2. The de-coating process has to be accurately controlled, as depletion of A3 factors leads to rapid turnover of pre-rRNAs (reviewed in ref. 1).

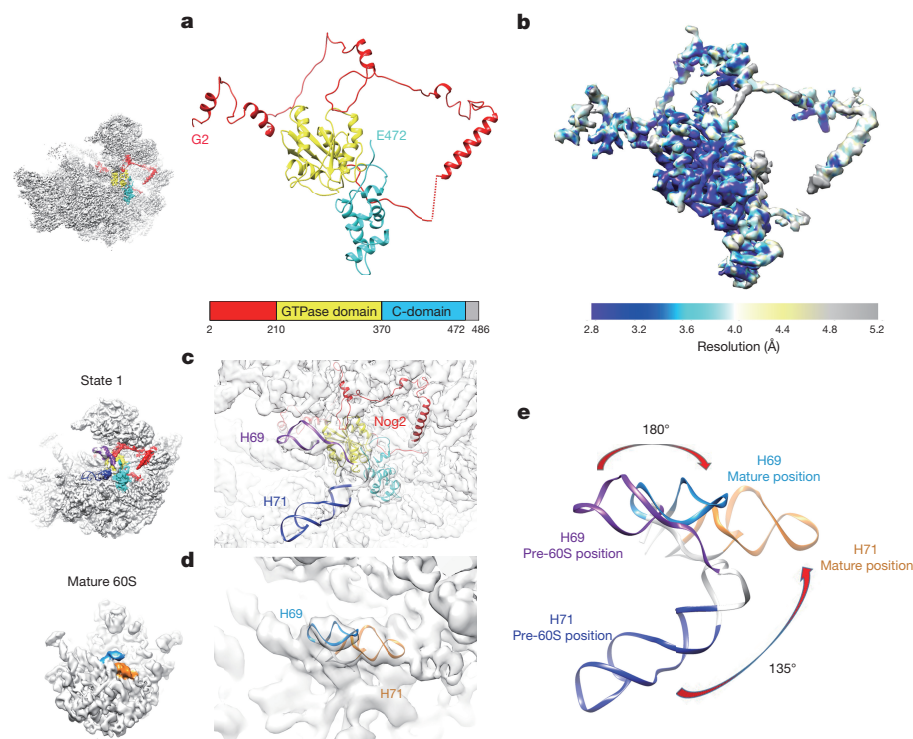


Figure 2 | Structure of Nog2 and its remodelling role in central helices H69–H71. **a**, Atomic structure of Nog2 (2–472 amino acids) with domains separately coloured, highlighting the N-terminal extension of Nog2. The orientation of Nog2 in the pre-60S particle is shown in the left thumbnail. **b**, Local resolution map of Nog2. Segmented Nog2 density map is coloured

according to the scale bar below. **c**, Zoom-in view of the H69–H71 region in the map of state 1, superimposed with atomic models of Nog2, H69 and H71. **d**, Same as **c**, but for the density map of the mature 60S subunit. **e**, Comparison of H69–H71 in the two density maps.

In the structure of state 1, the 5S RNP (the subcomplex of 5S rRNA, L5 and L11) is positioned almost 180°-rotated from its position in the mature subunit, as previously reported^{6,11}. Two central-protuberance-binding factors, Rpf2 and Rrs1, which anchor the 5S RNP to the pre-60S particles in an earlier stage^{13,29}, are apparently crucial to maintain this distinct conformation of the 5S RNP, as they provide

a support to the floating helical stem of the 5S RNP in the middle (Fig. 1a). In addition, rRNA helices of the central protuberance, stabilized by several interacting factors (Rpf2, Rrs1, Nsa2, Rsa4 and Nog2), display radically different conformations compared with those in the mature 60S subunit (Extended Data Fig. 8). Many of them are in completely topside-down or inside-out positions (Extended Data Fig. 8a).

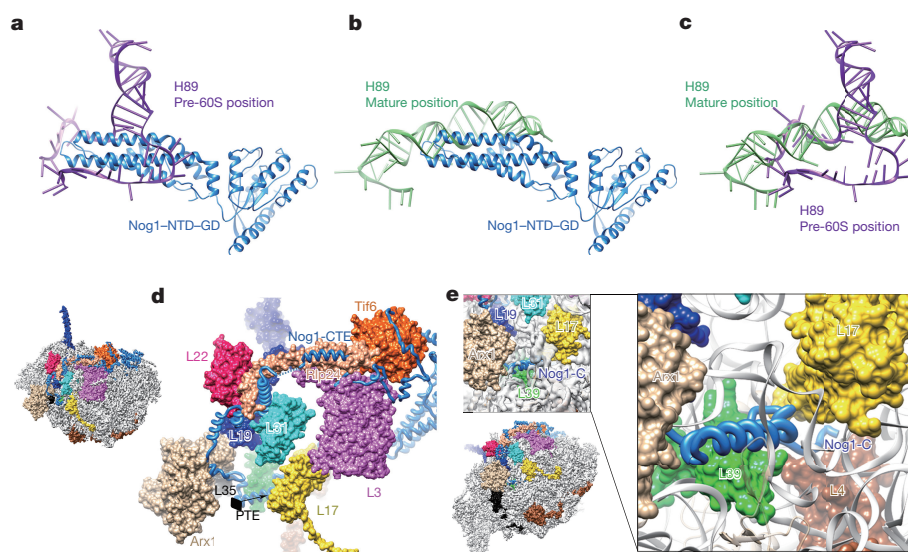


Figure 3 | Structure and binding partners of Nog1. **a**, The NTD of Nog1 inserts directly into the two strands of H89. GD, GTPase domain. **b**, Superimposition of the NTD of Nog1 with H89 in its mature conformation, displaying a steric clash in the terminal tip of the NTD of Nog1. **c**, Structural comparison of H89 in the pre-60S and mature conformations. **d**, The CTE of Nog1 interacts with multiple assembly

factors (Tif6, Rlp24, Arx1) and r-proteins (L3, L31, L22, L19, L35) in an arc region of the pre-60S particle. The overview is shown in the left thumbnail. The position and direction of polypeptide tunnel exit is denoted by a black diamond. **e**, The last C-terminal portion of Nog1 goes into the polypeptide tunnel and interacts with L4, L17, and L39 (see also Extended Data Fig. 6).

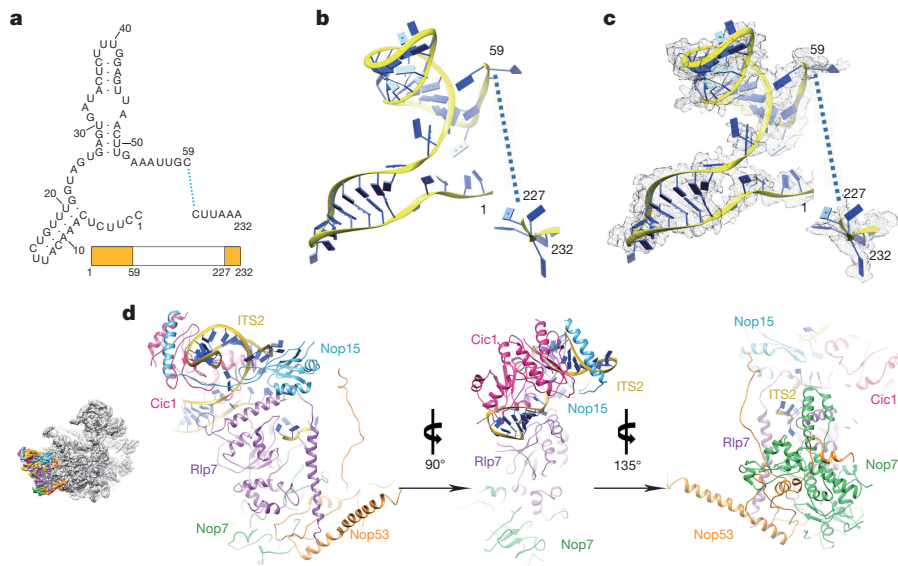


Figure 4 | Structure of ITS2 and associated factors. **a**, Secondary structure of the partial ITS2 rRNA sequences resolved in the map of state 1. **b**, Atomic model of the partial ITS2 rRNA. **c**, Same as **b**, with density map

The most dramatic change is that H80 is stretched into a single strand.

Comparison of the three states (1–3) indicates that the 5S RNP is in a different position in each state, reflecting snapshots of continuous rotational movement of the central protuberance (Extended Data Fig. 9a, b). In the structure of state 2, Rpf2 and Rrs1 are absent, and the 5S RNP has already rotated to a near-mature position, suggesting that removal of Rpf2 and Rrs1 is necessary for the rotation to occur. The structure of state 2 (~6.6 Å resolution) is very similar to that of the recently characterized Rix1–Rea1 particles¹⁴, containing five additional factors (Extended Data Fig. 9e–h): Sda1, Rix subcomplex (Ipi1, Rix1 and Ipi3), and Rea1. Sda1, with its characteristic HEAT (huntingtin, elongation factor 3, protein phosphatase 2A and lipid kinase TOR) repeat domain sandwiched between the L1 stalk and H38, pulls the L1 stalk into an inward position (Extended Data Fig. 9e, f). The Rix1 subcomplex sits above Sda1, contacting the gigantic remodelling ATPase Rea1 situated above the central protuberance¹⁴ (Extended Data Fig. 9g, h). Therefore, removal of Rpf2–Rrs1 might lead to binding of Sda1 and the Rix1 subcomplex, as well as Rea1 that subsequently releases Rsa4 (ref. 30). This last remodelling event enables further accommodation of the 5S RNP in the mature-like position observed in the structure of state 3. Notably, the stepwise conformational maturation of the 5S RNP is coordinated with sequential conformational changes of H38 in the three structures. Interestingly, repositioning of H38 from state 1 to state 2 involves the conformational change of the C terminus of Cgr1, from a bent helix to a straightened form (Extended Data Fig. 9c, d).

In summary, the rich atomic information presented in our structures provides a valuable resource to interpret and integrate a large body of existing genetic and biochemical data of eukaryotic ribosome assembly. In particular, it demonstrates potential diverse roles of assembly factors in late nuclear stages of large subunit assembly, and reveals unprecedented mechanistic details for two essential assembly GTPases, Nog1 and Nog2.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 16 November 2015; accepted 21 March 2016.

Published online 25 May 2016.

1. Woolford, J. L., Jr & Baserga, S. J. Ribosome biogenesis in the yeast *Saccharomyces cerevisiae*. *Genetics* **195**, 643–681 (2013).

superimposed. **d**, Three consecutively rotated views of ITS2 and associated factors. The orientation of the ITS2 subcomplex in the density map of state 1 is shown in the leftmost thumbnail.

2. Panse, V. G. & Johnson, A. W. Maturation of eukaryotic ribosomes: acquisition of functionality. *Trends Biochem. Sci.* **35**, 260–266 (2010).
3. Lebreton, A. *et al.* 60S ribosomal subunit assembly dynamics defined by semi-quantitative mass spectrometry of purified complexes. *Nucleic Acids Res.* **36**, 4988–4999 (2008).
4. McCann, K. L., Charette, J. M., Vincent, N. G. & Baserga, S. J. A protein interaction map of the LSU processome. *Genes Dev.* **29**, 862–875 (2015).
5. Saveanu, C. *et al.* Nog2p, a putative GTPase associated with pre-60S subunits and required for late 60S maturation steps. *EMBO J.* **20**, 6475–6484 (2001).
6. Leidig, C. *et al.* 60S ribosome biogenesis requires rotation of the 5S ribonucleoprotein particle. *Nature Commun.* **5**, 3491 (2014).
7. Saveanu, C. *et al.* Sequential protein association with nascent 60S ribosomal particles. *Mol. Cell. Biol.* **23**, 4449–4460 (2003).
8. Matsuo, Y. *et al.* Coupled GTPase and remodelling ATPase activities form a checkpoint for ribosome export. *Nature* **505**, 112–116 (2014).
9. Sengupta, J. *et al.* Characterization of the nuclear export adaptor protein Nmd3 in association with the 60S ribosomal subunit. *J. Cell Biol.* **189**, 1079–1086 (2010).
10. Greber, B. J., Boehringer, D., Montellese, C. & Ban, N. Cryo-EM structures of Arx1 and maturation factors Rei1 and Jji1 bound to the 60S ribosomal subunit. *Nature Struct. Mol. Biol.* **19**, 1228–1233 (2012).
11. Bradatsch, B. *et al.* Structure of the pre-60S ribosomal subunit with nuclear export factor Arx1 bound at the exit tunnel. *Nature Struct. Mol. Biol.* **19**, 1234–1241 (2012).
12. Kharde, S., Calviño, F. R., Gumiero, A., Wild, K. & Sinning, I. The structure of Rpf2–Rrs1 explains its role in ribosome biogenesis. *Nucleic Acids Res.* **43**, 7083–7095 (2015).
13. Madru, C. *et al.* Chaperoning 5S RNA assembly. *Genes Dev.* **29**, 1432–1446 (2015).
14. Barrio-Garcia, C. *et al.* Architecture of the Rix1–Rea1 checkpoint machinery during pre-60S-ribosome remodeling. *Nature Struct. Mol. Biol.* **23**, 37–44 (2016).
15. Klinge, S., Voigts-Hoffmann, F., Leibundgut, M., Arpagaus, S. & Ban, N. Crystal structure of the eukaryotic 60S ribosomal subunit in complex with initiation factor 6. *Science* **334**, 941–948 (2011).
16. Greber, B. J. *et al.* Insertion of the biogenesis factor Rei1 probes the ribosomal tunnel during 60S maturation. *Cell* **164**, 91–102 (2016).
17. Ben-Schem, A. *et al.* The structure of the eukaryotic ribosome at 3.0 Å resolution. *Science* **334**, 1524–1529 (2011).
18. Talkish, J., Zhang, J., Jakovljevic, J., Horsey, E. W. & Woolford, J. L. Jr. Hierarchical recruitment into nascent ribosomes of assembly factors required for 27SB pre-rRNA processing in *Saccharomyces cerevisiae*. *Nucleic Acids Res.* **40**, 8646–8661 (2012).
19. Dembowski, J. A., Kuo, B. & Woolford, J. L., Jr. Has1 regulates consecutive maturation and processing steps for assembly of 60S ribosomal subunits. *Nucleic Acids Res.* **41**, 7889–7904 (2013).
20. Feng, B. *et al.* Structural and functional insights into the mode of action of a universally conserved Obg GTPase. *PLoS Biol.* **12**, e1001866 (2014).
21. Pertsch, B. *et al.* Cytoplasmic recycling of 60S preribosomal factors depends on the AAA protein Drg1. *Mol. Cell. Biol.* **27**, 6581–6592 (2007).
22. Kappel, L. *et al.* Rlp24 activates the AAA-ATPase Drg1 to initiate cytoplasmic pre-60S maturation. *J. Cell Biol.* **199**, 771–782 (2012).
23. Granneman, S., Petfalski, E. & Tollervey, D. A cluster of ribosome synthesis factors regulate pre-rRNA folding and 5.8S rRNA maturation by the Rat1 exonuclease. *EMBO J.* **30**, 4006–4019 (2011).

24. Granato, D. C., Machado-Santelli, G. M. & Oliveira, C. C. Nop53p interacts with 5.8S rRNA co-transcriptionally, and regulates processing of pre-rRNA by the exosome. *FEBS J.* **275**, 4164–4178 (2008).
25. Thoms, M. *et al.* The exosome is recruited to RNA substrates through specific adaptor proteins. *Cell* **162**, 1029–1038 (2015).
26. Jakovljevic, J. *et al.* Ribosomal proteins L7 and L8 function in concert with six A₃ assembly factors to propagate assembly of domains I and II of 25S rRNA in yeast 60S ribosomal subunits. *RNA* **18**, 1805–1822 (2012).
27. Mitchell, P., Petfalski, E. & Tollervey, D. The 3' end of yeast 5.8S rRNA is generated by an exonuclease processing mechanism. *Genes Dev.* **10**, 502–513 (1996).
28. Thomson, E. & Tollervey, D. The final step in 5.8S rRNA processing is cytoplasmic in *Saccharomyces cerevisiae*. *Mol. Cell. Biol.* **30**, 976–984 (2010).
29. Zhang, J. *et al.* Assembly factors Rpf2 and Rrs1 recruit 5S rRNA and ribosomal proteins rpL5 and rpL11 into nascent ribosomes. *Genes Dev.* **21**, 2580–2592 (2007).
30. Ulbrich, C. *et al.* Mechanochemical removal of ribosome biogenesis factors from nascent 60S ribosomal subunits. *Cell* **138**, 911–922 (2009).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank the National Center for Protein Sciences (Beijing, China) for providing resource for cryo-EM data collection and

computation. We also thank members of Woolford laboratory for reading the manuscript. This work was supported by the Ministry of Science and Technology of China (2013CB910404 to N.G. and 2014CB849800 to M.-Q.D.), the National Natural Science Foundation of China (31422016 and 31470722 to N.G., and 21375010 to M.-Q.D.) and National Institutes of Health grant R01GM028301 (to J.L.W.).

Author Contributions N.G. and J.L.W. designed and directed experiments; B.K., H.B., M.G. and J.J. purified samples; D.T. and M.-Q.D. performed XL-MS; S.W. collected cryo-EM data (with J.L., Y.Y., Z.L., and C.M.), performed image processing (with Y.Z.), and analysed structures (with Y.K.). N.G., S.W. and K.Y. performed structural modelling. S.W., J.L.W. and N.G. wrote the paper.

Author Information The cryo-EM density maps of state 1 and state 2 have been deposited in the Electron Microscopy Data Bank under accession numbers EMD-6615 and EMD-6616, respectively. The atomic model of state 1 has been deposited in the Protein Data Bank (PDB) under accession number 3JCT. The XL-MS data have been deposited to the ProteomeXchange Consortium (<http://www.proteomexchange.org/>) with the dataset identifier PXD003736. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to J.L.W. (jw17@andrew.cmu.edu) or N.G. (ninggao@tsinghua.edu.cn).

METHODS

No statistical methods were used to predetermine sample size. The experiments were not randomized. The investigators were not blinded to allocation during experiments and outcome assessment.

Purification of Nog2-particles. Pre-ribosomes were purified by tandem affinity purification (TAP) with magnetic Dynabeads (Invitrogen) as explained previously³¹. TAP-tagged Nsa1, Nog2 and Nmd3 were used as baits to isolate ribosome assembly intermediates. The protein composition of the TAP-purified pre-ribosomes was determined by SDS-PAGE (4–10% Tris-glycine and 4–12% Bis-Tris, Invitrogen) followed by silver-staining³¹. Protein levels in each intermediate were assayed by western blotting analysis. Furthermore, the proteins associated with each intermediate were identified by mass spectrometry. Purified samples were sent to Penn State Hershey Core Research Facilities for trypsin digestion and matrix-assisted laser desorption/ionization–time of flight analysis. Results were analysed by Protein Pilot software and proteins identified with >99.9% confidence were used for further analysis.

XL-MS analysis. The Nog2-particles containing ~10 µg total proteins were incubated with BS³ or DSS at 1:1 (w/w) protein-to-cross-linker ratio at 25 °C for 1 h before the cross-linking reaction was quenched with 20 mM ammonium bicarbonate. Proteins were then precipitated with acetone, dissolved in 20 µl 8 M urea, 100 mM Tris, pH 8.5, and digested with trypsin at 37 °C overnight. Liquid chromatography–tandem mass spectrometry (LC-MS/MS) analyses of the digested samples were performed on an EASY-nLC 1000 system (Thermo Fisher Scientific) interfaced to a Q-Exactive mass spectrometer (Thermo Fisher Scientific). Peptides were separated on a 75 µm × 10 cm analytical column packed with 1.8 µm, 120 Å UHPLC-XB-C18 resin (Welch Materials) over a 110-min linear gradient made with buffer A (0.1% formic acid in HPLC-grade water) and buffer B (0.1% formic acid in HPLC-grade acetonitrile) as follows: 0–3 min, 0–5% B; 3–93 min, 5–30% B; 93–100 min, 30–80% B; 100–110 min, 80% B. The flow rate was set to 200 nl min⁻¹. The mass spectrometer was operated in data-dependent mode with one MS1 event at resolution 70,000 followed by ten HCD MS2 events at resolution 17,500. Dynamic exclusion time was set to 60 s. Precursors with a charge state of +1, +2 or unassigned were rejected. Three analytical replicates were performed for both BS³- and DSS-cross-linked samples. To identify proteins in the sample, we performed an additional LC-MS/MS analysis without rejecting precursors of +2 charge state, and the MS data were then searched against an *S. cerevisiae* protein database using ProLuCID³². After filtering the ProLuCID search results using DTASelect³³, 264 proteins were identified (false discovery rate for protein identity = 0.46%) and a database containing the sequences of these proteins was constructed for pLink search. Cross-linked peptides were identified using this database and the pLink software³⁴, and the results were filtered by requiring false discovery rate < 0.05, *E* value < 0.0001, and spectral count ≥ 2, which resulted in identification of 282 cross-linked peptide pairs (Supplementary Table 1). Results of XL-MS analysis, including information for peptide pair, statistical significance (*E* value), calculated mass, resolution (Δ mass) and mass accuracy (parts per million), are summarized in Supplementary Table 2. The XL-MS data have been deposited in the ProteomeXchange Consortium under data set identifier PXD003736, which contains one SEARCH file (pLink search result, false discovery rate < 0.05), seven PEAK files (ms2 files) and seven RAW files.

Cryo-EM data acquisition. Vitrified specimens were prepared by adding 4-µl samples of Nog2-particles at a concentration of ~150 nM to a glow-discharged holey carbon grid (Quantifoil R2/2) covered with a freshly made thin carbon film. Grids were blotted for 1 s and plunge-frozen into liquid ethane using an FEI Vitrobot Mark IV (4 °C and 100% humidity). Cryo-grids were transferred to an FEI Titan Krios electron microscope that was operating at 300 kV, and images were recorded using a K2 Summit direct electron detector (Gatan) in counting mode at a nominal magnification of ×22,500, corresponding to a pixel size of 1.32 Å at the object scale and with the defocus varying from –1.0 to –2.0 µm. All micrographs with K2 camera were collected using UCSF Image4 (developed by X. Li and Y. Cheng) under low-dose conditions. Each micrograph was dose-fractionated to 32 frames with a dose rate of ~8.2 counts per physical pixel per second for a total exposure time of 8 s. A fraction of micrographs were also recorded using Titan Krios (FEI) microscope operated at 300 kV under low-dose conditions with an FEI eagle 4k × 4k CCD camera, using an automated data collection software AutoEMation³⁵.

Image processing. Original image stacks were summed and corrected for drift and beam-induced motion at micrograph level using MOTIONCORR (developed by X. Li and Y. Cheng)³⁶. Programs of SPIDER³⁷ and EMAN2 (ref. 38) were used for micrograph screening, automatic particle picking and normalization. The contrast transfer function parameters of each micrograph were estimated by CTFFIND3 (ref. 39). All 2D and 3D classification and refinement were performed with RELION⁴⁰. Two-dimensional reference-free classification was applied to further screen particles (Extended Data Fig. 2a). At first, four batches of data were

collected (Extended Data Table 1) and processed separately following the same procedures. For each batch, particles were split into ten classes during the first round of 3D classification, with a map of the mature 60S ribosomal subunit (low-pass filtered to 60 Å) as the initial model. Based on the map features (the presence of ITS2 and the rotation of the 5S RNP), classes were combined and subjected the second and third rounds of 3D classification. Around 30% particles in the first four batches belong to state 1 (solid densities for ITS2 and the 5S RNP in a premature unrotated position). However, for the first four batches of data, particles displayed a strong orientation preference, which led to a noticeable distortion in the final density maps. Although the nominal resolutions of these maps were in the range of 3.8–4.5 Å, the distortion prevented accurate atomic modelling. To limit those strongly over-represented angular projections, SPIDER and RELION were used to balance the particles within different projection groups (by limiting the maximal number of particles for each projection group) during 3D refinement. Nevertheless, this additional procedure improved the map appearance to a certain extent, but could not completely eliminate the distortion in the final density maps. Another attempt was performed by combining the first four batches of data before 2D and 3D classification. All particles from the first four batches that belonged to state 1 were grouped together and subjected to 3D refinement with orientation-limiting procedure applied. However, the orientation preference still limited the high-resolution refinement and atomic modelling. Therefore, a series of optimizations in cryo-grid preparation were applied before the collection of the fifth data set, including elevated sample concentration, prolonged glow-discharge time, and reduced blotting time. As a result, there was no detectable orientation preference in the fifth data set (batch 8 in Extended Data Table 1a). For this batch of data, ~184,222 raw particles were picked from 833 micrographs for several rounds of reference-free 2D classification, yielding 143,707 good particles for 3D classification. A map of state 1 (low-pass filtered to 60 Å) was used as the initial reference for the 3D classification, which split the particles into eight classes (Extended Data Fig. 2b). One (A5) of the eight classes (8% of total particles) were discarded. Four of them belonged to state 1. The rest of these classes represented a series of intermediate structures. Another two batches of cryo-EM data were obtained (batches 9 and 10 in Extended Data Table 1a), which resulted in 304,296 particles for 3D classification into eight classes (B1–B8) (Extended Data Fig. 2b). Six of them (B3–B8) belonged to state 1, and as a result, they were combined for further high-resolution structural refinement. Comparison of state 1 structures from batch 8 and batch 9–10 indicates that the quality of last two batches of particles was slightly better, according to the density appearance of Cgr1 in the density map. Therefore, a homogeneous data subset (191,848 particles) for state 1 was obtained (B3–B8), from which a density map with 3.8-Å resolution (gold-standard Fourier shell correlation (FSC) 0.143 criteria) was constructed. To reduce the possible radiation damage to the particles, only frames 3–16 of each image stack were selected to generate a set of dose-reduced micrographs. A new set of particles were re-windowed from dose-reduced micrographs and subjected to 3D refinement, which improved the resolution to 3.6 Å. A soft-edged mask was then applied during final rounds of the high-resolution refinement, further improving the resolution to 3.46 Å. The final density map was corrected for the modulation transfer function of K2 detector, sharpened by applying a negative *B*-factor automatically estimated by post-processing program of RELION, and corrected for the soft-masked induced effects on FSC curves using high-resolution noise substitution⁴¹, resulting in a 3.08-Å density map for state 1. The local resolution map was estimated using ResMap⁴².

To further improve the density map of state 2, all non-state 1 particles from batches 8, 9 and 10 were combined (168,267 particles in total) and subjected to a round of 3D classification (Extended Data Fig. 2b) into eight classes. Around 6.5% of particles (10,900) belonged to state 2 (C8), and refinement of these particles rendered a final density map at a nominal resolution of 6.6 Å.

Model building and refinement. Crystal structure of the yeast 80S ribosome (PDB accession number 3U5D)¹⁷ was used as the initial template for rRNA modelling. The models of the rRNAs (25S, 5.8S) were docked into the density map manually using UCSF Chimera⁴³. The 5S rRNA was separately fitted into its density by rigid-body docking. The crystal structure of the 25S rRNA was compared with that of the Arx1-TAP pre-60S structure (PDB accession number 3J64)⁶ in the density map, and fragments of nucleotides 995–1054, 2244–2318, 2615–2771 and 2789–2804 of the crystal structure were cut out and fitted into our density map. After the initial fitting, the entire chains of rRNAs were manually checked and adjusted with COOT⁴⁴.

For modelling of ITS2 RNA, secondary structures were predicted using RNAfold⁴⁵ and drawn using RnaViz⁴⁶. Atomic modelling of ITS2 RNA was performed with COOT, started with a poly-adenine model, followed by sequence replacement.

For r-protein modelling, structures of individual proteins from the crystal structure of yeast 80S ribosome (PDB accession number 3U5E)¹⁷ were separately fitted

into the density map using Chimera. Except for L10, L24, L29, L40, L41 and L42, which are absent in the density map of Nog2-particles, chains of the remaining r-proteins were manually adjusted using COOT. Structures of L5 and L11 were first docked into the density map in a subcomplex with the 5S rRNA, followed by a similar manual adjustment in COOT.

For modelling of biogenesis factors, the sequences of all factors according to the result of mass spectrometry (Extended Data Fig. 1b) were subjected to 2D and 3D structure prediction, using PSIPRED⁴⁷ and I-TASSER⁴⁸, respectively. Initial fitting of biogenesis factors, such as Nog2, Nop15, Rlp7, Nop7 and Cic1, was guided by previous biochemical data and our XL-MS data, and was confirmed by high agreement of secondary structural features between the predicted 3D models and the density map. Specifically, for each factor, the five 3D models predicted by I-TASSER were aligned in PYMOL⁴⁹, and the common structural motifs were selected and used for rigid body fitting in Chimera. Taking Nog2 as an example, the GTPase domain (residues 207–369) and the C-terminal domain (residues 373–486) were first separately fitted into Chimera, followed by manual adjustment of main chains and side chains in COOT. Linker building and further extension of chains in both the N- and C-directions were done manually in COOT. Information from secondary structural prediction was used to aid main-chain tracing. In many cases, poly-alanine models were first built, and sequence assignments were aided by well-resolved bulky residues such as Phe, Tyr, Trp and Arg. As for factors Nog1, Rlp24, Rsa4, Arx1 and Mrt4, the initial positions of them were taken from a previous low-resolution cryo-EM studies^{6,10}, followed by extensive model rebuilding in COOT. In particular, main-chain tracing of the C-terminal extension of Nog1 was done completely manually. For factors Rpf2 and Rrs1, their *S. cerevisiae* models were generated using CHAINSAW⁵⁰ in the CCP4 suite⁵¹ with the crystal structures of *Aspergillus nidulans* Rpf2 and Rrs1 (PDB accession number 4XD9 and 5BY8)^{12,52} as templates. The crystal structure of the yeast Tif6 (PDB accession number 1G62)⁵³ was docked into the density map to provide an initial position. The modelling of the factors in the ITS2 region was largely facilitated by our XL-MS data, as some portions of these factors were not resolved in the map. The modelling of Nsa2 was facilitated by the crystal structure of Rsa4 in complex with Nsa2 peptide (PDB accession number 4WJV)⁵⁴, which provided an anchor point for the N- and C-terminal halves of Nsa2 during atomic modelling.

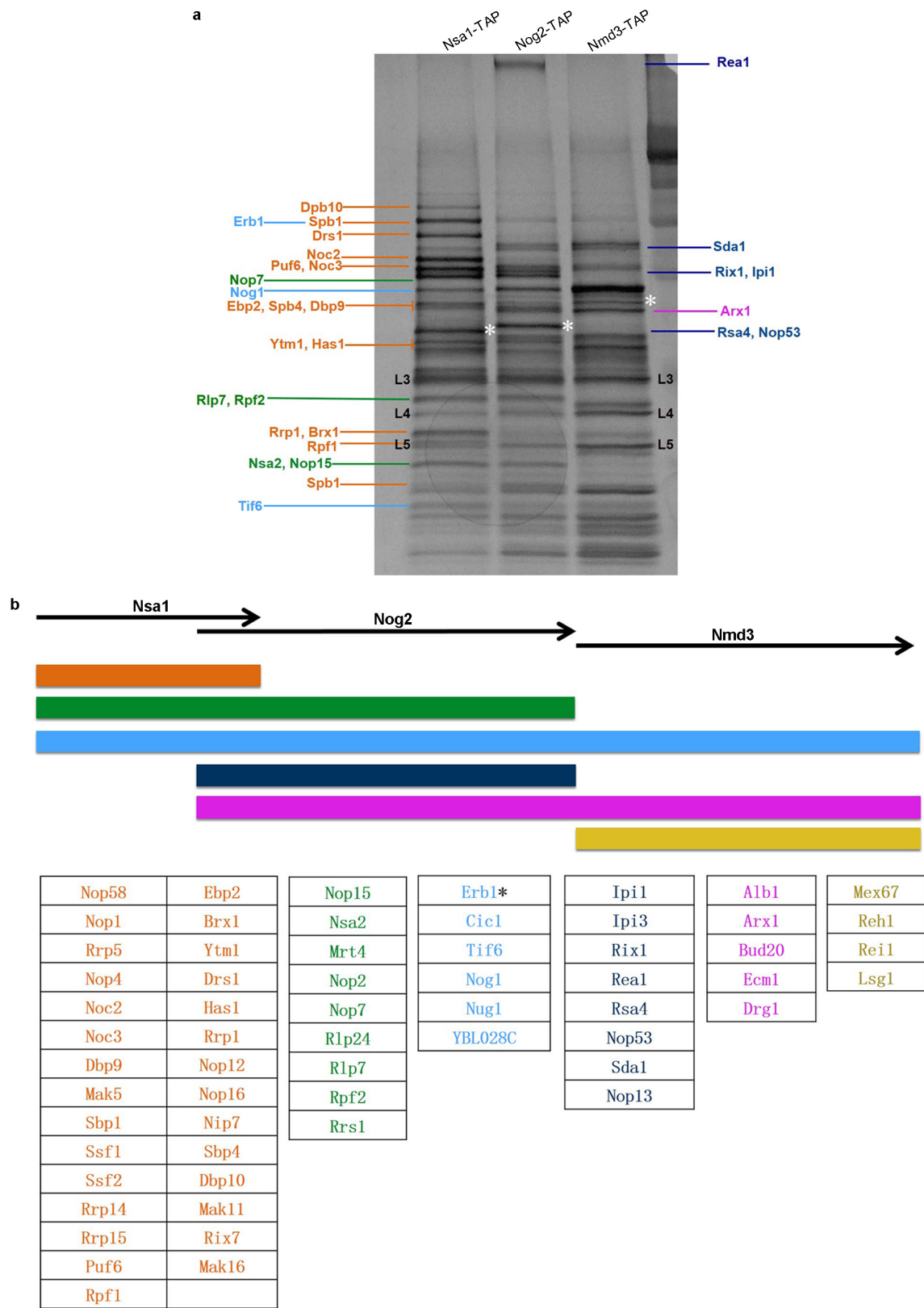
Docking of Sda1 and Rea1 in the density map of state 2 was facilitated by the recent cryo-EM study of Rix1–Rea1 particles¹⁴. The models of Sda1 and Rea1 (PDB accession number 5FL8)¹⁴ were fitted into our density map of state 2 as rigid bodies (Extended Data Fig. 9e–h).

The atomic model of state 1 containing ribosomal proteins, rRNAs and assembly factors was refined against the density map first by real-space refinement (phenix.real_space_refine)⁵⁵ in PHENIX⁵⁶ with secondary structure and geometry constraints applied. After refinement, alternating rounds of manual model adjustment using COOT and model refinement using PHENIX were applied. A final round of model refinement was done in Fourier space using REFMAC⁵⁷ with secondary structure, base pair and planarity restraints applied, according to previously established protocols⁵⁸. To avoid overfitting, different weights of the density map for refinement were tested. Cross-validation against overfitting was performed following the procedures previously described^{58,59}. The atom positions of the atomic model were randomly displaced by 0.5 Å before the model was refined against a map reconstructed from half of the data (named Half1 map) produced by RELION during the last iteration of high-resolution structural refinement. And two FSC curves were calculated on the basis of refined model: one was FSC_{work} (model versus Half1 map) and the other was FSC_{test} (model versus Half2 map). In addition, another FSC curve was calculated for the comparison of refined model with final density map. Comparison of FSC_{test} and FSC_{work} curves showed no large separation between them, indicating the final atomic model was not overfitted. Statistics of final model was evaluated using MolProbity⁶⁰ (Extended Data Table 1b).

Of the 282 cross-linked peptide pairs identified in the XL-MS data, the distances of 151 lysine pairs could be calculated from the model of state 1. Ninety-four per cent of them (142) agree with the model with the Cα–Cα distances ≤ 24 Å between

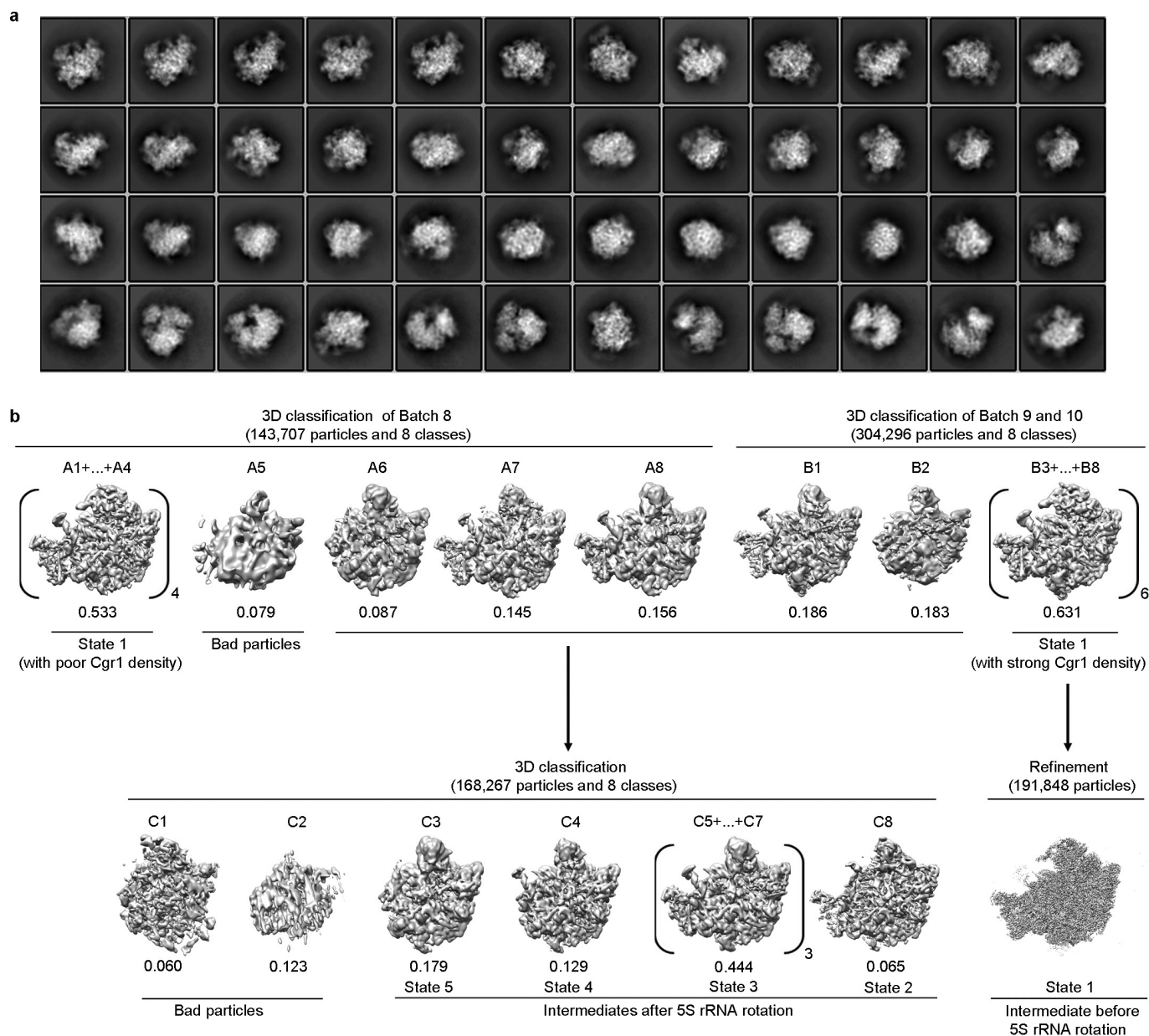
two cross-linked lysine residues. Among the incompatible nine pairs, five of them are with the Cα–Cα distances ≤ 30 Å.

- Sahasranaman, A. *et al.* Assembly of *Saccharomyces cerevisiae* 60S ribosomal subunits: role of factors required for 27S pre-rRNA processing. *EMBO J.* **30**, 4020–4032 (2011).
- Xu, T. *et al.* ProLuCID, a fast and sensitive tandem mass spectra-based protein identification program. *Mol. Cell. Proteomics* **5**, S174 (2006).
- Tabb, D. L., McDonald, W. H. & Yates, J. R., III. DTASelect and Contrast: tools for assembling and comparing protein identifications from shotgun proteomics. *J. Proteome Res.* **1**, 21–26 (2002).
- Yang, B. *et al.* Identification of cross-linked peptides from complex samples. *Nature Methods* **9**, 904–906 (2012).
- Lei, J. & Frank, J. Automated acquisition of cryo-electron micrographs for single particle reconstruction on an FEI Tecnai electron microscope. *J. Struct. Biol.* **150**, 69–80 (2005).
- Li, X. *et al.* Electron counting and beam-induced motion correction enable near-atomic-resolution single-particle cryo-EM. *Nature Methods* **10**, 584–590 (2013).
- Shaikh, T. R. *et al.* SPIDER image processing for single-particle reconstruction of biological macromolecules from electron micrographs. *Nature Protocols* **3**, 1941–1974 (2008).
- Tang, G. *et al.* EMAN2: an extensible image processing suite for electron microscopy. *J. Struct. Biol.* **157**, 38–46 (2007).
- Mindell, J. A. & Grigorieff, N. Accurate determination of local defocus and specimen tilt in electron microscopy. *J. Struct. Biol.* **142**, 334–347 (2003).
- Scheres, S. H. RELION: implementation of a Bayesian approach to cryo-EM structure determination. *J. Struct. Biol.* **180**, 519–530 (2012).
- Chen, S. *et al.* High-resolution noise substitution to measure overfitting and validate resolution in 3D structure determination by single particle electron cryomicroscopy. *Ultramicroscopy* **135**, 24–35 (2013).
- Kucukelbir, A., Sigworth, F. J. & Tagare, H. D. Quantifying the local resolution of cryo-EM density maps. *Nature Methods* **11**, 63–65 (2014).
- Pettersen, E. F. *et al.* UCSF Chimera—a visualization system for exploratory research and analysis. *J. Comput. Chem.* **25**, 1605–1612 (2004).
- Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. Features and development of Coot. *Acta Crystallogr. D* **66**, 486–501 (2010).
- Lorenz, R. *et al.* ViennaRNA Package 2.0. *Algorithms Mol. Biol.* **6**, 26 (2011).
- De Rijk, P., Wuyts, J. & De Wachter, R. RnaViz 2: an improved representation of RNA secondary structure. *Bioinformatics* **19**, 299–300 (2003).
- Buchan, D. W., Minneci, F., Nugent, T. C., Bryson, K. & Jones, D. T. Scalable web services for the PSIPRED Protein Analysis Workbench. *Nucleic Acids Res.* **41**, W349–W357 (2013).
- Yang, J. *et al.* The I-TASSER Suite: protein structure and function prediction. *Nature Methods* **12**, 7–8 (2015).
- Schrodinger, LLC. The PyMOL molecular graphics system, version 1.3r1 (2010).
- Stein, N. CHAINSAW: a program for mutating pdb files used as templates in molecular replacement. *J. Appl. Cryst.* **41**, 641–643 (2008).
- Winn, M. D. *et al.* Overview of the CCP4 suite and current developments. *Acta Crystallogr. D* **67**, 235–242 (2011).
- Asano, N. *et al.* Structural and functional analysis of the Rpf2–Rrs1 complex in ribosome biogenesis. *Nucleic Acids Res.* **43**, 4746–4757 (2015).
- Groft, C. M., Beckmann, R., Sali, A. & Burley, S. K. Crystal structures of ribosome anti-association factor IF6. *Nature Struct. Biol.* **7**, 1156–1164 (2000).
- Baßler, J. *et al.* A network of assembly factors is involved in remodeling rRNA elements during preribosome maturation. *J. Cell Biol.* **207**, 481–498 (2014).
- Afonine, P. V. *et al.* Towards automated crystallographic structure refinement with phenix.refine. *Acta Crystallogr. D* **68**, 352–367 (2012).
- Adams, P. D. *et al.* PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr. D* **66**, 213–221 (2010).
- Murshudov, G. N., Vagin, A. A. & Dodson, E. J. Refinement of macromolecular structures by the maximum-likelihood method. *Acta Crystallogr. D* **53**, 240–255 (1997).
- Amunts, A. *et al.* Structure of the yeast mitochondrial large ribosomal subunit. *Science* **343**, 1485–1489 (2014).
- Fernández, I. S., Bai, X. C., Murshudov, G., Scheres, S. H. & Ramakrishnan, V. Initiation of translation by cricket paralysis virus IRES requires its translocation in the ribosome. *Cell* **157**, 823–831 (2014).
- Chen, V. B. *et al.* MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallogr. D* **66**, 12–21 (2010).

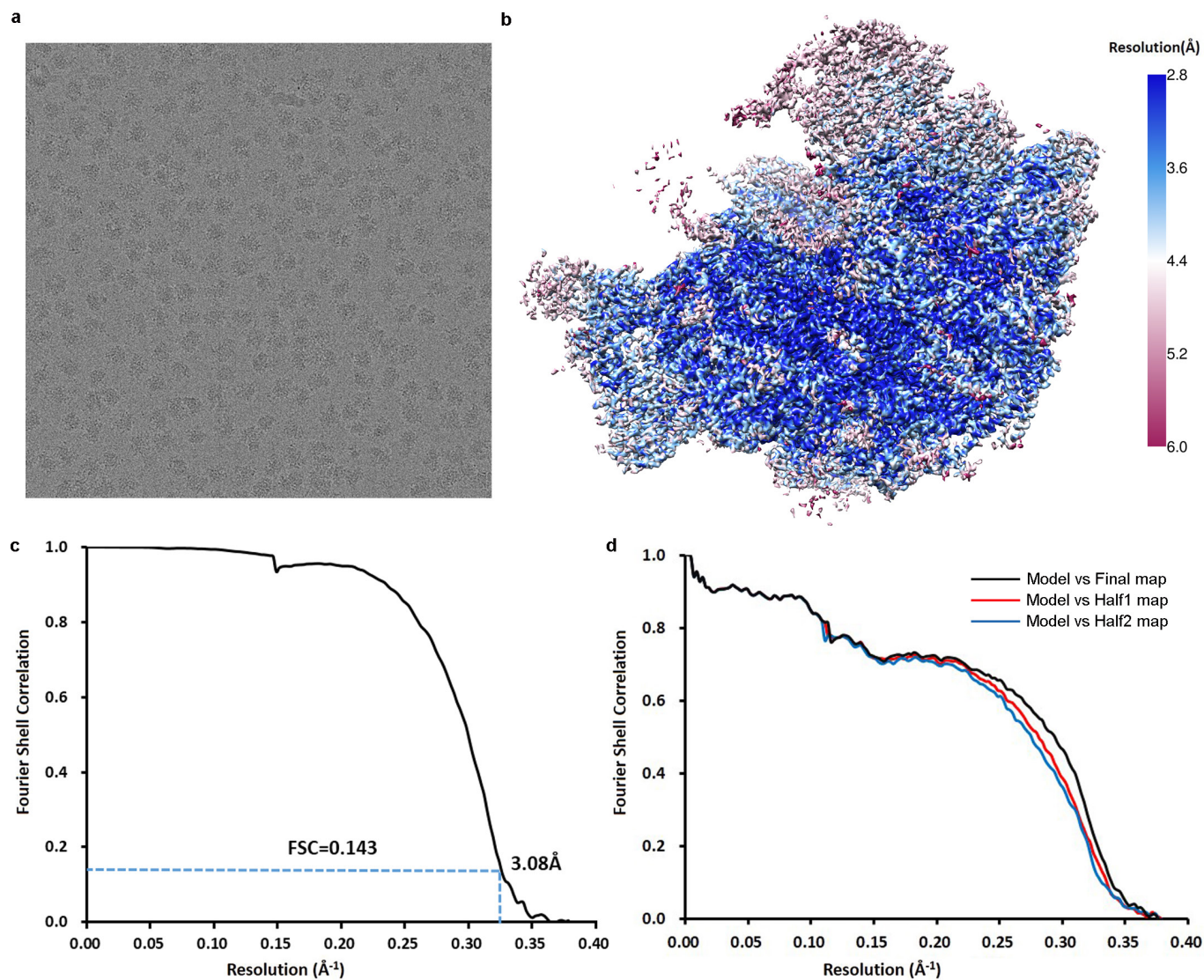


Extended Data Figure 1 | Compositional analysis of Nsa1, Nog2 and Nmd3 particles. **a**, Mostly non-overlapping assembly factors Nsa1, Nog2 and Nmd3 were used to purify sequential ribosome assembly intermediates. Proteins identified by mass spectrometry analysis were marked on the gel. Orange coloured proteins are only present in Nsa1-TAP particles, green coloured proteins are present both in Nsa1-TAP and in Nog2-TAP particles, light blue coloured proteins are present in all three purified particles to varying levels, dark blue coloured proteins are present only in Nog2-particles, pink coloured proteins are present both in Nog2- and Nmd3-particles in varying levels and yellow coloured proteins

are present only in Nmd3-particles. TAP-tagged proteins are indicated by white asterisks. For gel source data, see Supplementary Fig. 1. **b**, The lifetimes of mostly non-overlapping ribosome assembly intermediates containing assembly factors Nsa1, Nog2 and Nmd3 are indicated. Assembly factors identified in each of Nsa1-TAP, Nog2-TAP and Nmd3-TAP associated samples were colour coded. The colour scheme is identical to that used in **a**. *Even though this protein was identified in all three intermediates, its levels decreased more than sevenfold from Nsa1-TAP particles to Nog2-TAP particles.

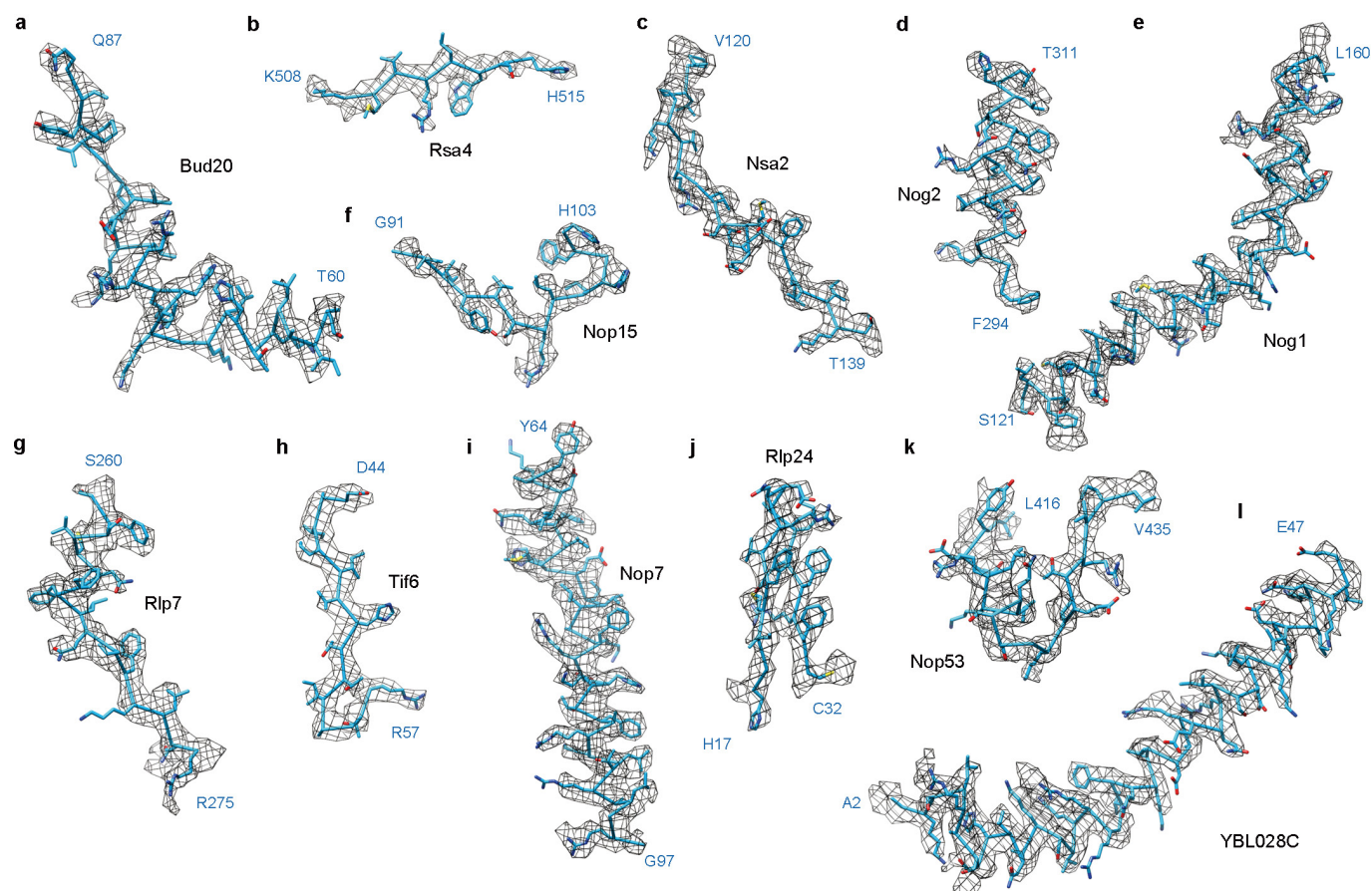


Extended Data Figure 2 | Cryo-EM data processing of Nog2-particles. **a**, Representative 2D class averages of Nog2-particles. **b**, A flow-chart for 3D classification of Nog2-particles (data batch 8–10, see Methods for details).

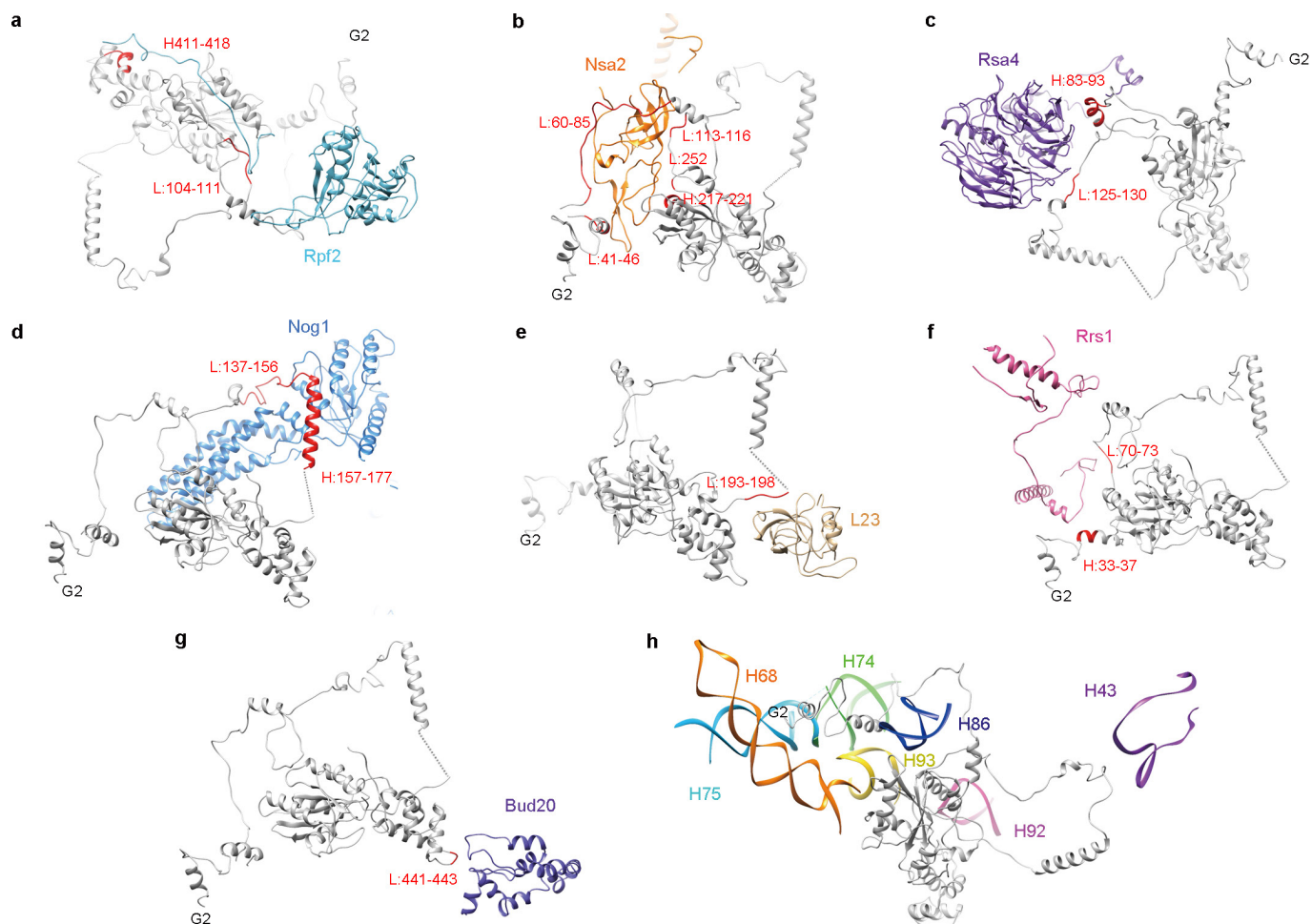


Extended Data Figure 3 | Resolution estimation and model validation.
a, Representative micrograph of Nog2-particles. **b**, Local resolution map of the final density map of state 1. **c**, FSC curve for the final density map (state 1). The nominal resolution is 3.08 Å estimated using the gold-standard (FSC = 0.143) criterion. **d**, Atomic model cross-validation.

Three FSC curves were calculated between the refined model (against Half1 map) and the final map (black), between the refined model with Half1 map (FSC_{work}, red), and between the refined model with Half2 map (FSC_{test}, blue) (see Methods for details).

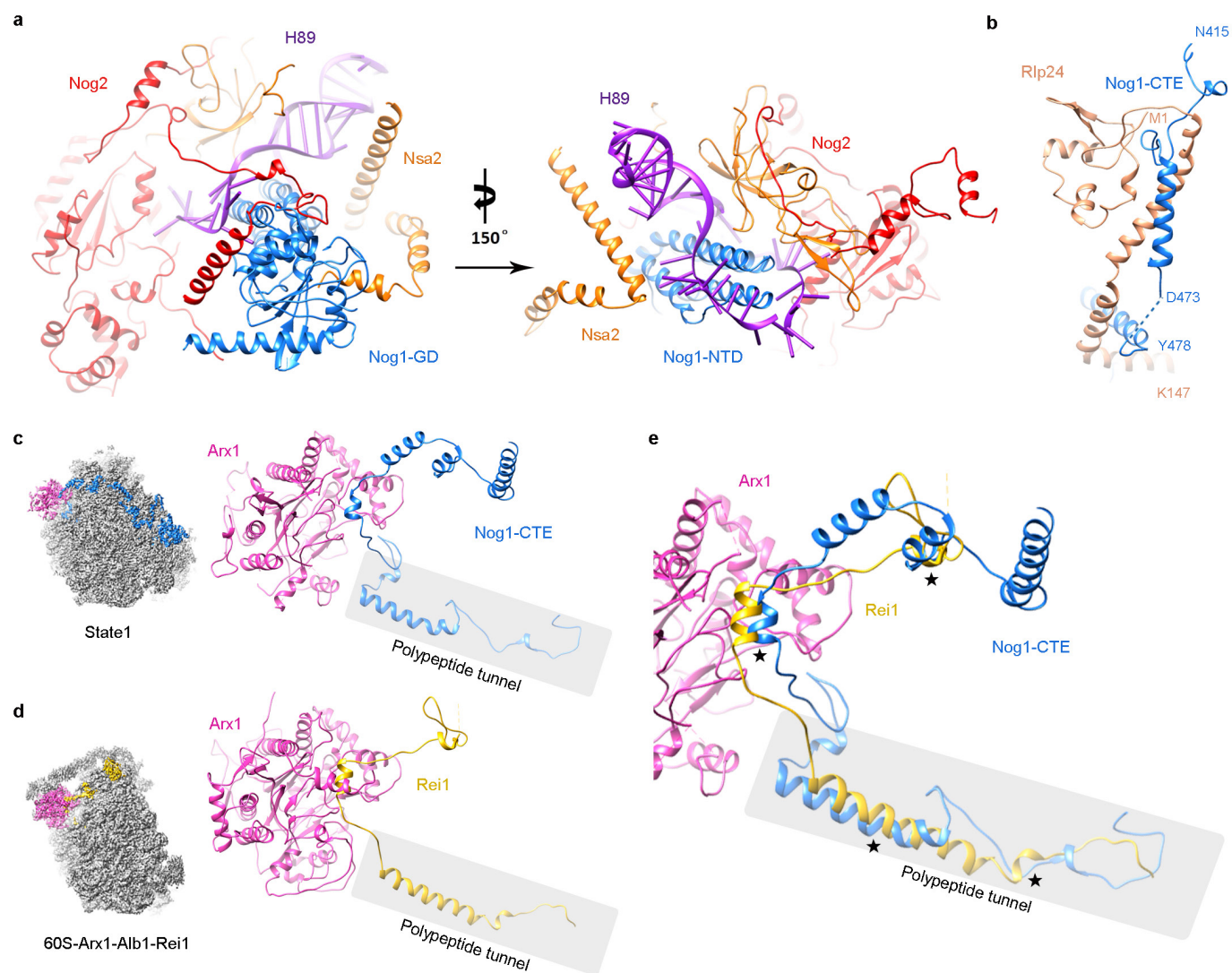


Extended Data Figure 4 | Local densities of representative regions for different assembly factors. **a–l**, Cryo-EM densities of representative regions of assembly factors, superimposed with respective atomic models.



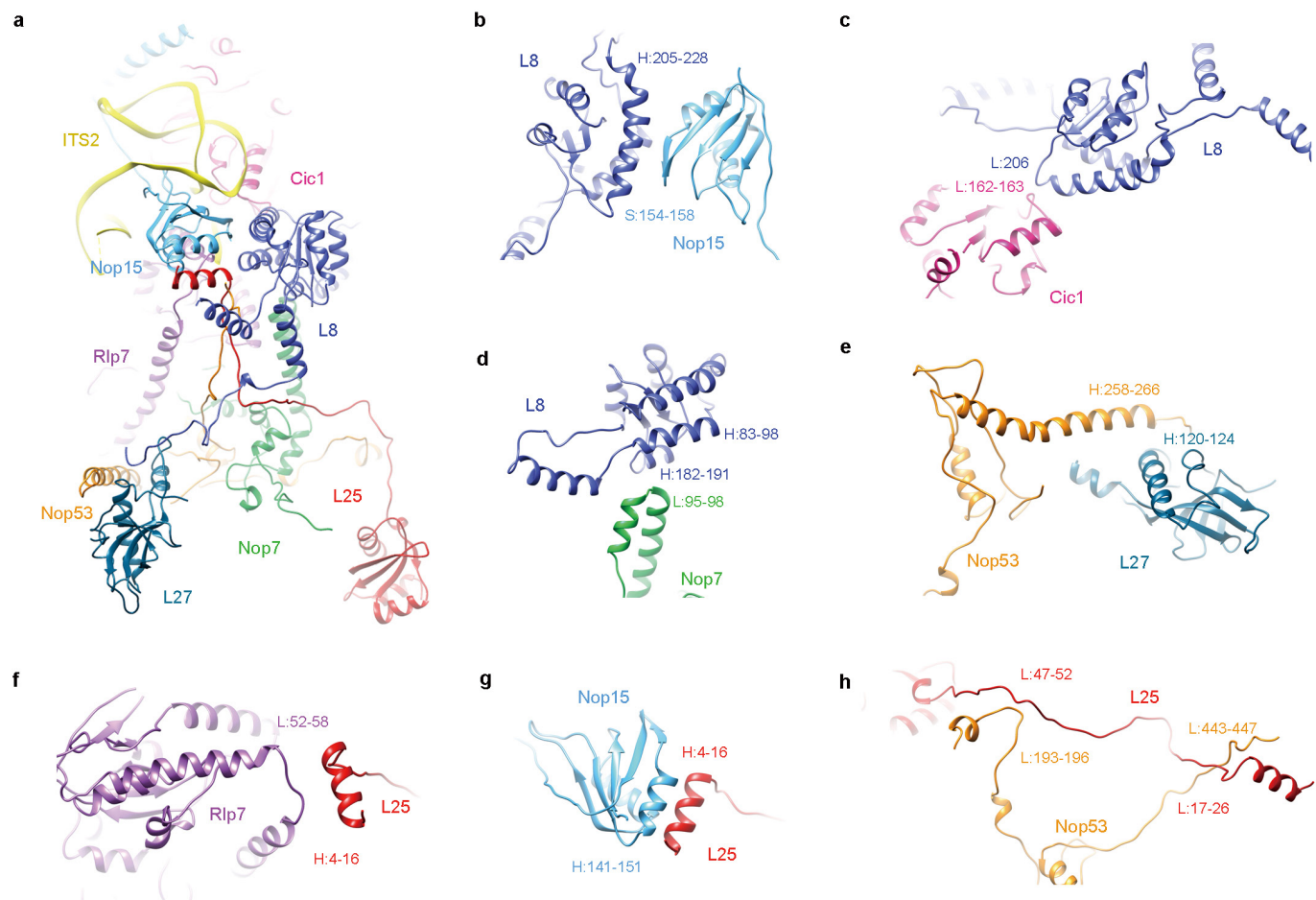
Extended Data Figure 5 | Interaction network of Nog2 in the pre-60S particle. a–g. Pairwise illustration of binding partners of Nog2 in the pre-60S particle. Residues of Nog2 involved in atomic contacts are coloured red with residue numbers labelled. H and L denote helix and

loop, respectively. **h.** Interactions between rRNA components (H43, H68, H74, H75, H86, H92, H93) and Nog2. For clarification, H69 and H71 are not shown. The N terminus of Nog2 is located in a helical junction composed of H68, H74, H75 and H93.



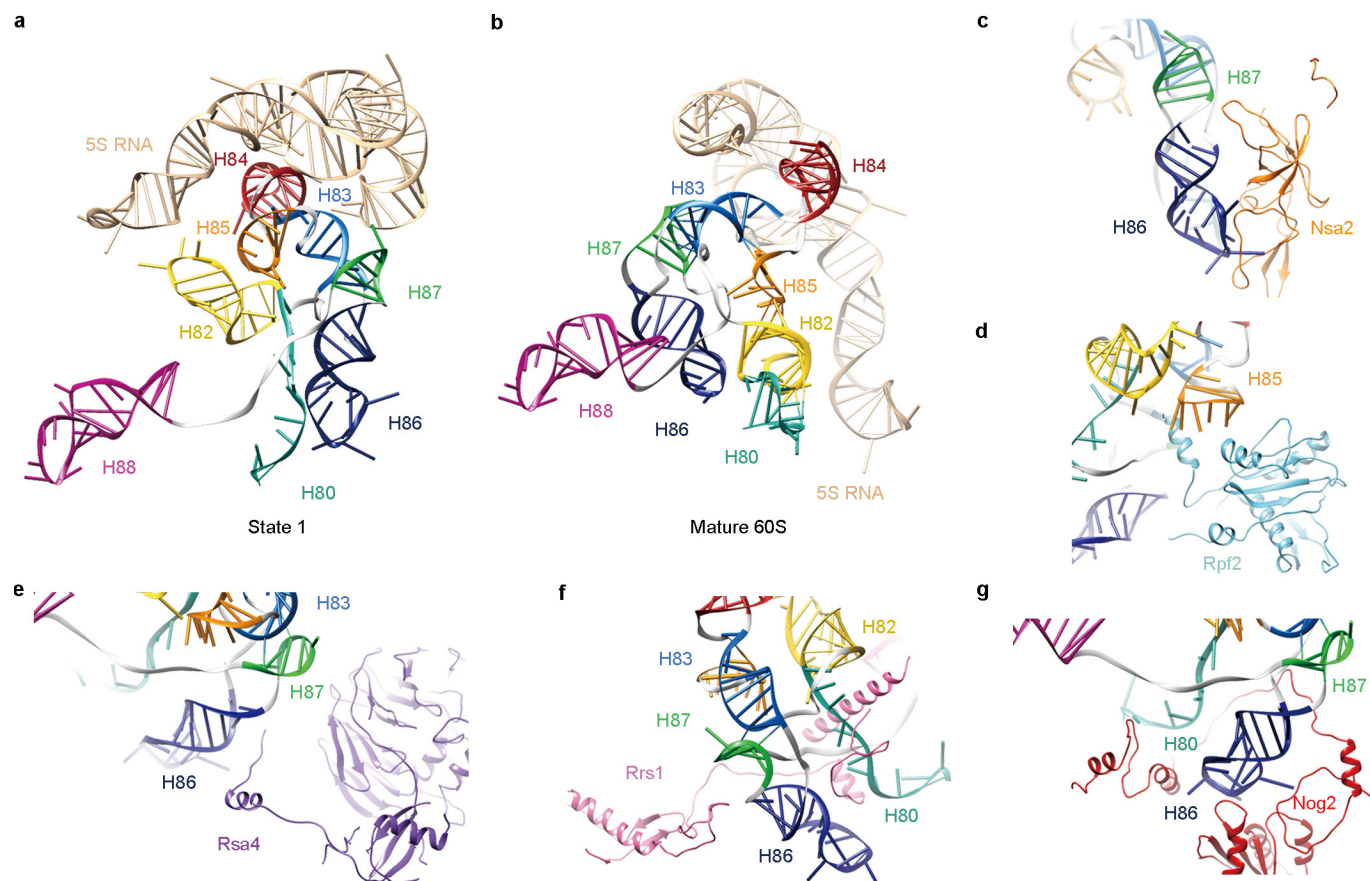
Extended Data Figure 6 | The NTD of Nog1 interacts with Nsa2 and Nog2. a, Nsa2, Nog2 and Nog1 collectively stabilize H89 in a distinct conformation. Nog1 interacts with Nog2 and Nsa2 through its GTPase domain and NTD, respectively. **b**, The CTE of Nog1 interlocks with Rlp24 by wrapping around a long helix at the C-terminal end of Rlp24 (see also Fig. 3). **c**, **d**, Comparison of the CTE of Nog1 and the CTE of Rei1

in the polypeptide tunnel. Atomic models of state 1 (**c**) and 60S-Arx1-Alb1-Rei1 (**d**) (PDB accession number 5APN)¹⁶ are aligned using the 60S subunit. For clarification, only Arx1, Nog1 and Rei1 are shown. **e**, Superimposition of **c** and **d**. Four major places of steric clash between Rei1 and Nog1 are marked by asterisks.



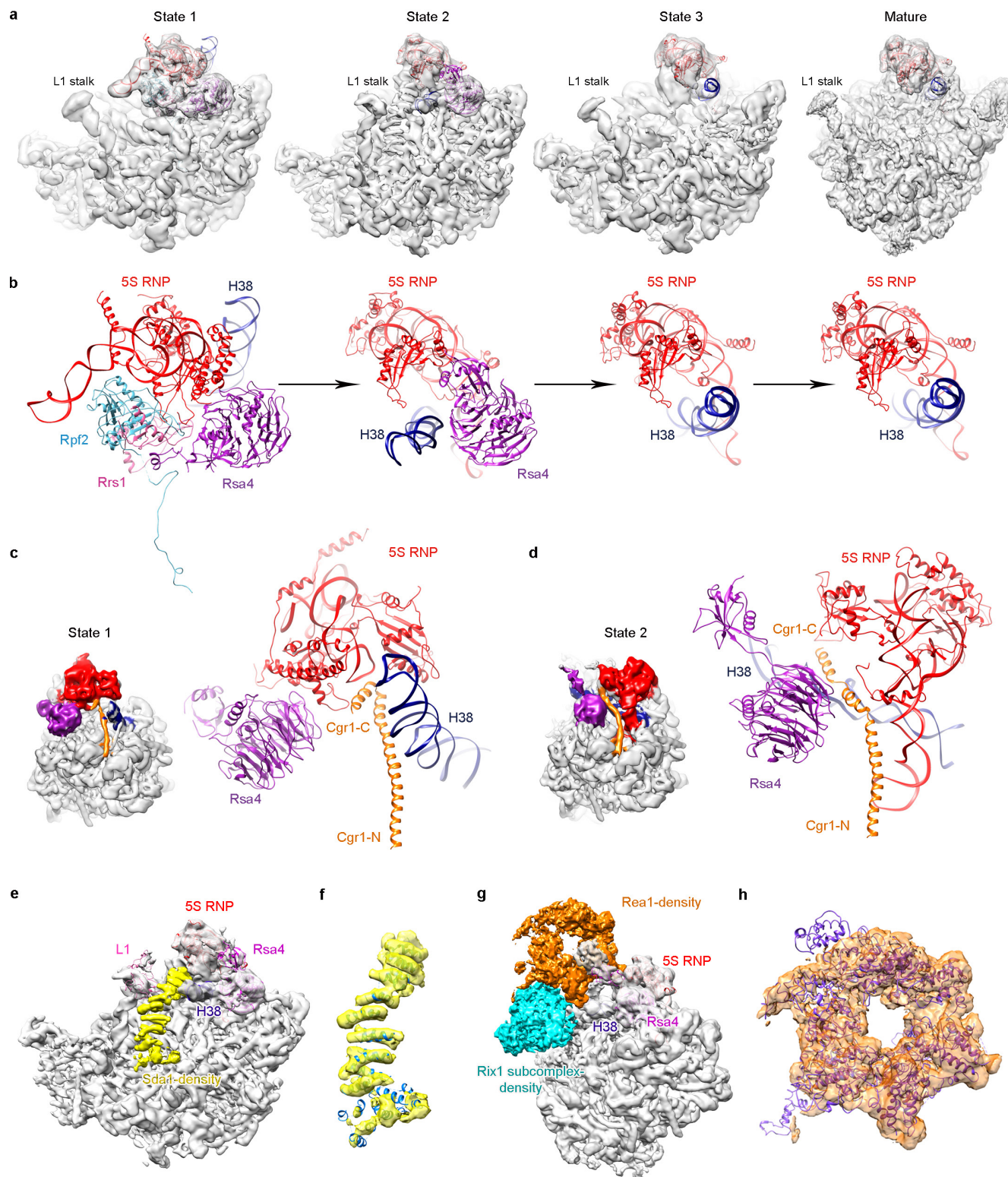
Extended Data Figure 7 | Mutual interactions between factors and r-proteins in the ITS2 subcomplex. **a**, An overall view of the ITS2 subcomplex. **b–d**, L8 interacts with three factors: Nop15 (**b**), Cic1 (**c**) and Nop7 (**d**). **e**, L25 interacts with Nop53. **f–h**, L25 interacts with Rlp7 (**f**),

Nop15 (**g**) and Nop53 (**h**). Residues involved in atomic interaction sites are labelled with sequence numbers. H, L, S denote helix, loop and strand of respective structures.



Extended Data Figure 8 | Restructuring of rRNA helices in the central protuberance region by Nsa2, Rpf2, Rsa4, Rrs1 and Nog2.
a, Conformation of rRNA helices from the central protuberance (H80, H82-H88, 5S rRNA) in the pre-60S particle (state 1). **b**, Same as **a**, but

for the mature 60S subunit. The mature 60S subunit was aligned to state 1 structure globally. **c–g**, Pairwise interactions between the central protuberance helices and factors are shown in separate panels.



Extended Data Figure 9 | Structures of different assembly states of the pre-60S ribosomal particles. **a**, Cryo-EM density maps of three premature states (1–3) and the mature state are displayed in transparent surface representation, superimposed with models of the 5S RNA, H38 and associated central-protubance-binding factors. **b**, Zoom-in views of the central protubance regions in **a**. For clarification, only atomic models are shown. Comparison of these four states indicates that the 5S RNP rotates to a near-mature state (state 2) after Rpf2–Rrs1 leave, and further release of Rsa4 in state 3 results in a ‘mature-like’ conformation for the 5S RNP. H38 from these four states is in a series of continuous changes coupled with the 5S RNP conformational maturation. **c**, **d**, Spatial relationship of the 5S RNP, H38, Rsa4 and Cgr1 in state 1 (**c**) and state 2 (**d**).

Note that repositioning of H38 from state 1 to state 2 is coupled with a dramatic conformational change on the C-terminal end of Cgr1. **e–h**, Additional assembly factors identified in the density map of state 2. One piece of additional density between H38 and L1 contains a characteristic HEAT repeat, which contacts the L1 stalk in an inward position (**e**). The atomic model of Sda1 (PDB accession number 5FL8)¹⁴ fits well with the segmented density (**f**). For clarification, densities immediately above Sda1 are not shown in **e** and **f**. A large piece of additional density in the map of state 2, composed of the Rix1 subcomplex and Rea1 (**g**, **h**). The density assignment was facilitated by the cryo-EM structure of Rix1–Rea1 particles¹⁴. Superimposition of the atomic model of Rea1 (PDB accession number 5FL8)¹⁴ with the segmented density map of Rea1 (**h**).

Extended Data Table 1 | Statistics of data collection, structural refinement and model validation

Batches	Electron Microscope	Camera	Micrographs (Original micrographs)	Particles for 2D classification	Particles for 3D classification
1	F20	US4000	381(966)	76,323	19,253
2	Titan Krios	Eagle	3,184(4,701)	154,785	133,455
3	Titan Krios	K2	1,017(1,136)	90,888	35,096
4	Titan Krios	K2	1,497(1,579)	200,292	100,956
5	Titan Krios	K2	997(1,002)	128,515	54,574
6	Titan Krios	K2	1,114(1,114)	139,309	54,257
7	Titan Krios	K2	1,014(1,016)	134,937	50,334
8	Titan Krios	K2	833(852)	184,222	143,707
9	Titan Krios	K2	901(901)	225,167	146,349
10	Titan Krios	K2	1,019(1019)	248,518	157,947

Data Collection

EM equipment	FEI Titan krios
Voltage (kV)	300
Detector	Gatan K2
Particles	191,848
Pixel size (Å)	1.32
Defocus range (µm)	1.0-2.0
Electron dose (e ⁻ /Å ²)	50 (32 frames)/22 (frame 3-16)

Model composition

Peptide chains	54
Protein residues	13,982
RNA chains	3
RNA bases	3,446

Refinement

Resolution (Å)	3.08
Map sharpening B-factor (Å ²)	-65
R factor	0.3040
Fourier Shell Correlation	0.7814

Rms deviations

Bonds (Å)	0.0054
Angles(°)	0.9687

Validation (proteins)

Molprobity score	2.43 (96 th percentile)
Clashscore, all atoms	3.44 (100 th percentile)
Good rotamers (%)	80.87

Ramachandran plot

Favored (%)	88.14
Outliers (%)	3.46

Validation (RNA)

Correct sugar puckers (%)	97.16
Good backbone conformations (%)	71.60

CORRIGENDUM

doi:10.1038/nature17420

Corrigendum: Observation of polar vortices in oxide superlattices

A. K. Yadav, C. T. Nelson, S. L. Hsu, Z. Hong, J. D. Clarkson,
C. M. Schlepütz, A. R. Damodaran, P. Shafer, E. Arenholz,
L. R. Dedon, D. Chen, A. Vishwanath, A. M. Minor, L. Q. Chen,
J. F. Scott, L. W. Martin & R. Ramesh

Nature **530**, 198–201 (2016); doi:10.1038/nature16463

In this Letter, the surname of author Christian M. Schlepütz was incorrectly spelled “Schlepüetz”. This has been corrected in the online versions of the paper.

CORRIGENDUM

doi:10.1038/nature16997

Corrigendum: Signalling thresholds and negative B-cell selection in acute lymphoblastic leukaemia

Zhengshan Chen, Seyedmehdi Shojaee, Maike Buchner, Huimin Geng, Jae Woong Lee, Lars Klemm, Björn Titz, Thomas G. Graeber, Eugene Park, Ying Xim Tan, Anne Satterthwaite, Elisabeth Paietta, Stephen P. Hunger, Cheryl L. Willman, Ari Melnick, Mignon L. Loh, Jae U. Jung, John E. Coligan, Silvia Bolland, Tak W. Mak, Andre Limnander, Hassan Jumaa, Michael Reth, Arthur Weiss, Clifford A. Lowell & Markus Müschen

Nature **521**, 357–361 (2015); doi:10.1038/nature14231

In Extended Data Fig. 3b of this Letter, 52 flow cytometry dot plots with double stainings for CD19 and ITIM-bearing receptors (PECAM1, LAIR1, CD300A and BTLA) were shown for 13 samples. The CD19-CD300A staining for sample ICN1 was inadvertently replaced with CD19-CD300A staining for sample PDX2. The Supplementary Information for this Corrigendum contains the corrected Extended Data Fig. 3b (showing the correct dot plot for sample ICN1). Our conclusions are not affected.

Supplementary Information is available in the online version of the Corrigendum.

ERRATUM

doi:10.1038/nature17622

Erratum: Epithelial tricellular junctions act as interphase cell shape sensors to orient mitosis

Floris Bosveld, Olga Markova, Boris Guirao, Charlotte Martin, Zhimin Wang, Anaëlle Pierre, Maria Balakireva, Isabelle Gaugue, Anna Ainslie, Nicolas Christophorou, David K. Lubensky, Nicolas Minc & Yohanns Bellaïche

Nature **530**, 495–498 (2016); doi:10.1038/nature16970

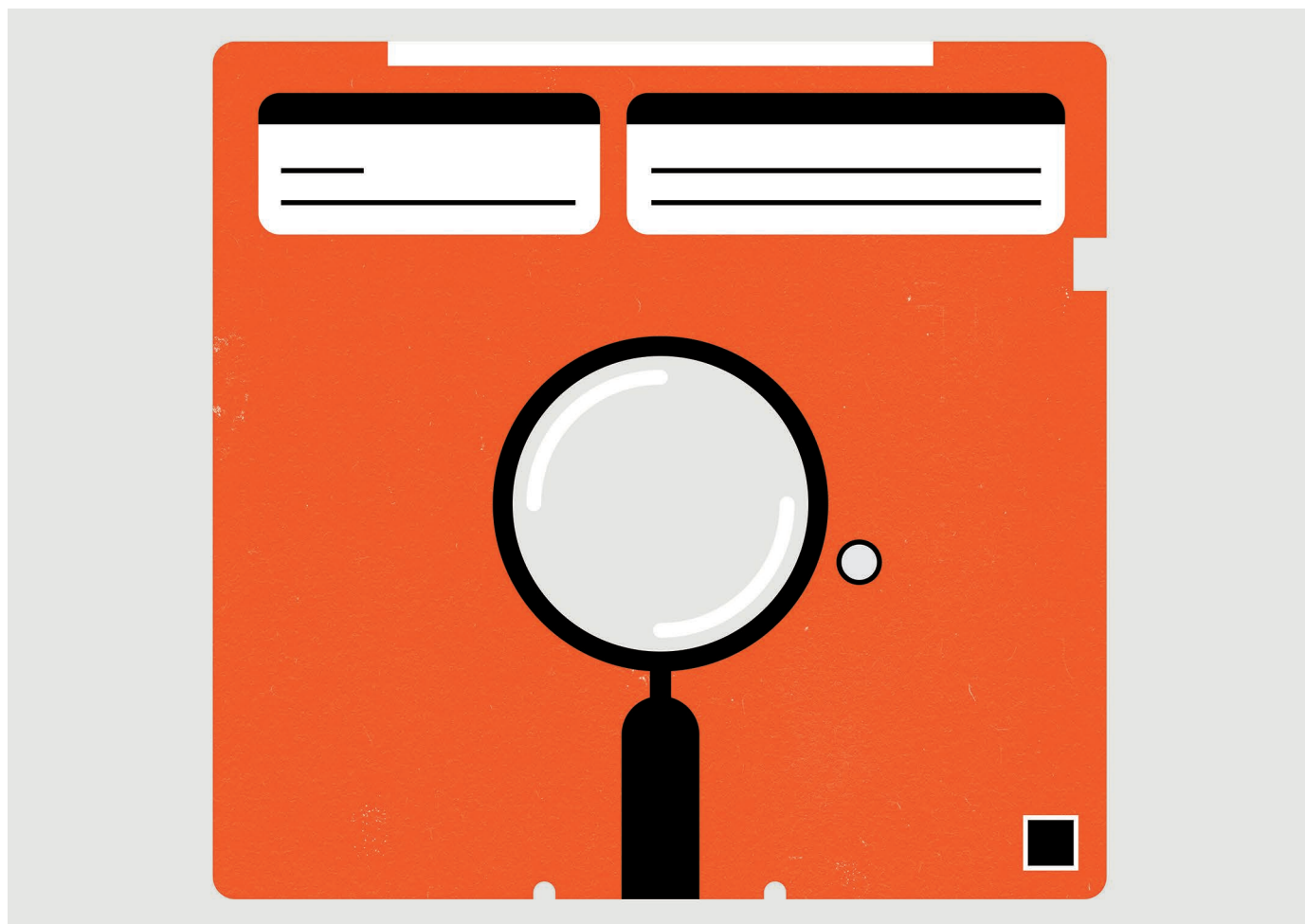
In Fig. 1e of this Letter, the *y*-axis label incorrectly read: ‘GFP-Mud intensity (TCJs per junction)’ instead of ‘GFP-Mud intensity (at TCJs relative to septate junctions)’. This has been corrected in the online versions of the paper.

TOOLBOX

DIGITAL FORENSICS IN THE LIBRARY

Archivists are borrowing and adapting techniques used in criminal investigations to access data and files created in now-obsolete systems.

ILLUSTRATION BY THE PROJECT TWINS



BY MARK WOLVERTON

When archivists at California's Stanford University received the collected papers of the late palaeontologist Stephen Jay Gould in 2004, they knew right away they had a problem. Many of the 'papers' were actually on computer disks of various kinds, in the form of 52 megabytes of data spread across more than 1,100 files — all from long-outdated systems.

"It was a large collection, as you can

imagine," says Michael Olson, service manager for the Born Digital/Forensics Lab at Stanford University Libraries. "He used a lot of early word processing for his writing, lots of disks and diskettes in different formats."

After considerable effort the Stanford archivists did get Gould's papers into order — first by finding hardware that could read the obsolete disks, and then by deciphering what they found there. "We had some challenges finding old applications to figure out what word processor he used, that sort of thing,"

says Olson.

The Gould papers were an early indication of an issue that's been rapidly worsening: four decades after the personal-computer revolution brought word processing and number crunching to the desktop, the first generation of early adopters is retiring or dying. So how do archivists recover and preserve what's left behind?

"People around the world have information stored on disks that are less readable with every passing day," says Christopher ►

► Lee, a researcher in the School of Information and Library Science at the University of North Carolina (UNC) in Chapel Hill. “This includes floppies, Zip disks, CDs, DVDs, flash drives, hard drives and a variety of other media.” Many files can be accessed only with long-obsolete hardware, and all are subject to physical deterioration that will ultimately make them unreadable by any means. By now, many libraries, archives and museums have accumulated shelves full of such material, stashed away in the hope that if it’s ever needed, somebody, somewhere will be able to figure out how to access it.

DIGITAL INSPIRATION

Increasingly, archivists are finding inspiration in the field of digital forensics: the art of extracting evidence about illicit activity from computer drives, smartphones, tablets or even GPS devices. “It turned out that law-enforcement and computer-security people were dealing with essentially the same problems of stabilizing and recovering data from digital media,” says Matthew Kirschenbaum at the University of Maryland in College Park. And many of their solutions were directly applicable to the archivists’ needs.

In law enforcement, for example, a top priority is to preserve material in its original form. This is often harder than it sounds: almost anything done on a computer, even something as innocuous as plugging in a USB drive, leaves a faint digital trace. So digital-forensics practitioners have developed techniques for creating an artefact-free ‘disk image’ that duplicates everything, down to the unused and hidden disk space. They can then preserve the integrity of the original for evidentiary purposes in court while doing all their forensic analysis on a perfect copy.

Institutions working to decipher collections have the same need, although in their case, the object is to maintain the provenance of the original for future researchers. Creating forensic copies of the data was a relatively fringe idea 8 or 10 years ago, Lee says. “It’s now quite common in library and archive settings.”

Unfortunately for archivists, however, disk imaging is usually done through commercial software packages such as the Forensic Toolkit made by Access Data in Lindon, Utah, or by EnCase, which is developed by Guidance

Software in Pasadena, California. Because these packages are designed for criminal investigators, they include tools for file carving (assembling complete files from fragmentary data); cracking passwords; accessing encrypted files; advanced searching; and generating reports for use in court — tasks that tend to be less important for archival purposes. These packages also come with licensing costs in the thousands of dollars, which would strain the budget of many collecting institutions.

So in 2011, Lee and his colleagues launched BitCurator, a platform designed for the archival field, with funding from the Andrew W. Mellon Foundation, and with continued support from a consortium that currently encompasses 25 member institutions, including Harvard University, the Massachusetts Institute of Technology, Stanford University, Emory University and the British Library. BitCurator has the advantage of being open source and freely available for download (wiki.bitcurator.net). “It’s a combination of third party open-source tools and our own

work,” says Kam Woods, a research scientist at UNC’s School of Information and Library Science and co-principal investigator with Lee on the project. On the basis of the turnout at training sessions and other BitCurator events, Lee estimates that several dozen institutions now use the package actively, and several hundred more use it at least occasionally.

BitCurator not only handles disk imaging, but a number of other issues that criminal investigators don’t have to worry about. One example is redaction: editing out confidential material before publication. That’s an alien concept in the criminal investigations, says Olson. “Why would you ever want to redact evidence from a case? But from an archival or library standpoint, you wouldn’t want to make somebody’s health records available.” So BitCurator has to have methods for access control that don’t really exist in the forensics field.

Another speciality of BitCurator is its

ability to read long-outdated disks — an essential tool for archivists who are faced with stacks of old floppies or even reels of magnetic tape. Although digital-forensics investigators usually deal with newer-generation systems, their techniques can still be quite useful for recovery, says Lee. “Taking a forensic approach, you can still create a safe copy of the data, even if you don’t know what the file system is or you can’t read it,” he says. “As long as you can attach a drive and get the bits off of it, you can create an image.” Archivists can then experiment on different ways to retrieve the files, safe in the knowledge that the original is not in danger.

Some advantages to the forensics-based approach transcend technical considerations, says Olson. With the Gould archives, for example, “you can get timestamps from different word-processing files to see how he actually wrote something, a particular order that he wrote, a way that he edited. That’s really nifty if you’re a researcher that wants to know how his mind worked.”

SEARCH AND RESCUE

The same techniques can be used for other purposes besides archiving. At Stanford, Olson’s lab is increasingly helping faculty members and students who need to access work that was born on now-outdated computer systems. “I had a graduate student about a year ago that came to us with an astrophysics data set on a Zip disk,” he says. “It was something that their professor had created, that they weren’t able to read and needed to get to because it was part of their research. And nobody had really shepherd that to a new modern system.” The library was able to help the student do just that.

Another recent example is Stanford’s long-running ME310 engineering course, which had a server full of design studies, presentation slides and videos that students had completed over the years as part of their graduate work. “The people running the programme wanted to preserve all the data from these projects,” says Olson, “but they needed help to recover the data, organize it and also get permission from the students to actually make this available.”

Data are already being lost to science at a rapid rate. One study, for example, found that as little as 20% of data for ecology papers published in the early 1990s is still available (T. H. Vines *et al. Curr. Biol.* **6**, 94–97; 2014). Co-author Tim Vines, who now runs a peer-review service called Axios Review in Vancouver, Canada, says that the best way for scientists to preserve their data for future generations is to upload it into library-maintained archives or open online repositories, such as Dryad or Figshare.

“Putting it into the hands of an organization committed to preserving it is far better than putting it on a shelf”, he says. ■

“People around the world have information stored on disks that are less readable with every passing day.”

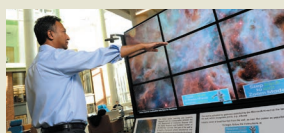


STAY CURRENT

- Scientists write their papers online — together go.nature.com/mcwlow
- Researchers move their reference libraries to the clouds go.nature.com/i5lhwp
- Computer scientists clean up ‘link rot’ go.nature.com/b9agxg

IN THE NEWS

Librarians and researchers are racing to cope with a flood of open data go.nature.com/r5k6tw



CAREERS

NANOTECHNOLOGY An Arab scientist in Israel builds non-invasive medical tests **p.143**

SWITCH How to move from natural science to data science go.nature.com/oomnml

NATUREJOBS For the latest career listings and advice www.naturejobs.com

BERT VOGELSTEIN



Postdoctoral researchers at the Ludwig Center at Johns Hopkins in Baltimore, Maryland, participate in an annual Halloween contest.

TEAM BUILDING

Morale boosters

You can keep spirits up when the research doldrums hit.

BY KENDALL POWELL

Every early-career scientist has been there: six months pass with no good news to report at a lab meeting. You can't move on to the next phase of your project because you still need to complete a particular experiment or analysis. Dread overtakes you at the thought of facing yet another week of try and try again.

Senior researchers know that this is the norm, not the exception. Making cutting-edge discoveries means that you may figuratively bash your head against a brick wall for many months before any true breakthroughs happen.

But even the most resilient junior researchers can get depressed and frustrated when weeks of experiments leave them empty-handed. The best group leaders know how to keep morale flying high in the face of the research doldrums

(see 'Let them eat cake'). They use one-on-one meetings and progress reports to keep the wheels rolling, and lab outings, group-bonding activities and silly contests to keep the 'fun' in functional labs. Many also have strategies to 'normalize' failure in their labs, so that researchers won't hide or grow listless when projects aren't working.

Principal investigators (PIs) and lab heads who want to maintain high spirits — and high productivity — in their labs need to keep in mind that morale flows from the top down, say veteran group leaders. It is crucial for PIs to foster community building in their lab to give team members a sense of belonging and to establish a support network that will see junior members through the roughest spots.

"The success of people in the lab at certain times can elevate the success of everyone," says

Jeff Karp, who leads a bioengineering group at Brigham and Women's Hospital in Cambridge, Massachusetts. He and other PIs make sure to acknowledge and celebrate successes publicly when they do arrive. "It's important to have in place a high morale because it helps bring out the best in everybody," he says.

BREAKOUT TIME

Stephen Royle, group leader at the University of Warwick, UK, knows that he is not necessarily the first person to whom his team members will turn when things aren't going well on a project in his cell-biology lab. So he tries to create and nurture a high level of trust and camaraderie in his group to establish a safety net that won't let any one person's woes reach crisis level.

His group goes on regular lunch and laser-tag outings together, and he's also dreamt up ►

► some friendly competitions to keep things light in the laboratory. Using the free pens, T-shirts and other trinkets gathered at scientific conferences for prizes, he runs a Lab Quiz event akin to pub quizzes. Lab members compete on trivia questions about the university and its host city of Coventry, and on little-known facts about clathrin, the lab's favourite molecule.

Royle's lab members chart lab records — who can get the maximum yield from a DNA preparation, say — on a whiteboard, and one wall hosts the 'Western blot Hall of Shame' with horribly smeared protein samples. Such a display is a great equalizer and morale-builder, Royle says — it shows students and junior lab members that even the most accomplished senior teammate with a stellar publication record can run experiments that deliver rubbish results. He also likes to tell his team about past blunders of big names — such as how Nobel laureate Roderick MacKinnon broke the electrode of a pH meter during his first days of undergraduate research.

It's important to "de-pathologize" the idea of failure in science, says astronomer Keivan Stassun, senior associate dean for graduate education and research at Vanderbilt University's College of Arts and Science in Nashville, Tennessee. Instead, you teach students that failures are more akin to having writers' block, a normal part of the process that everyone runs into. It is imperative, he adds, that no one feel isolated in their struggles.

With that in mind, many lab heads say that they try to flatten hierarchies and promote a connected, cooperative environment. One hierarchy-shattering activity that Karp uses is a three-minute presentation competition. Everyone in the lab gives a quick talk on any topic they choose — these have ranged from the best hamburger joints in the region to historical figures from India — and then gets feedback. The group votes on the best presentation and the best critique, and each winner receives a useful electronic gizmo, such as a slide pointer.

To promote collegiality, some PIs take their team-building ventures outdoors. When cell biologist Anne Straube was asked to organize a

LET THEM EAT CAKE

Good ways to wind down at the end of the week

At the end of a dispiriting week in the lab, it doesn't hurt to have something to look forward to as a morale lift. For PhD student Alice Bachmann, that is Cake Club — a weekly meeting of researchers in her building that is centred around the simple act of eating scrumptious cake. "Always on Fridays. It's a good way to finish the week with a high sugar level and high level of happiness," says Bachmann, who studies cell biology at the University of Warwick, UK.

"I want to keep people sane," says Anne Straube, Bachmann's adviser. "After they've spent a week in the microscope room counting dots on the computer screen, they need to break out of this." Although she doesn't quite remember how Cake Club started, she thinks that excelling at recipes in the kitchen helps lab members to hone their experimental skills, too.

Creations have included cakes in the shape of a brain, a green fluorescent

cheesecake and a cake decorated like the page of a PhD thesis. Members are also encouraged to experiment and bring their failures (which are still edible, after all).

'Vino', a social hour with refreshments, ends the week at the Vanderbilt Initiative in Data-Intensive Astrophysics at Vanderbilt University in Nashville, Tennessee. The gathering often includes a toast to research successes — getting a paper accepted, a fellowship awarded or grants funded.

The VINO room contains no projector or whiteboard, so no one can inadvertently slip too far into shop talk, and the informal chatting makes advisers and other professors more approachable for trainees. "We've been collectively working hard all week. Hopefully, we've had some glimmers of success, but we mostly all experience a whole lot of slog and tribulation," says Vanderbilt astronomer Keivan Stassun. "And we all deserve a beer." **K.P.**

social activity for her department's retreat, she warned them that it could get muddy — she's a fan of orienteering, which can mean running through ditches or alongside creeks. She designed a course on her campus at the University of Warwick: teams of 3 had to find the locations of 45 picture clues. Once people have collective fun and get to know each other, it makes it easier to ask for advice or a reagent, she says. "It breaks down the hierarchies and lets them be less serious about things."

Running two lab groups — one in Uppsala, Sweden, and one in Milan, Italy — means that cancer researcher Elisabetta Dejana must work doubly hard to ensure cohesiveness. She holds lab meetings once a week over Skype, and in January, organized a retreat for everyone to meet and gather in Milan. She hosted both groups at her home for pizza, gelato and wine, and took them on hikes.

The informal gathering tightened group bonding and eased the tension over competition between the groups. "They are exchanging messages and mice and cells," she says. "Now, they understand that working together will help everyone." When peer reviewers send back a research paper with calls for new experiments, she divides the work up among the group, which makes the revisions go more quickly and adds energy to the team, she says.

CAREFUL SCREENING

Creating a strong sense of community often starts with the recruitment process. Cancer researcher Bert Vogelstein accepts only applicants whom he thinks will be able to hold up emotionally and psychologically through the

challenges of bench science. During interviews, he probes candidates' mindsets: asking them how long they expect to work on their project, for example, and whether they have failed at anything before. "If they say, 'No', that's a conversation ender — they are not being honest with themselves," Vogelstein says. He also phones their previous mentors to glean detailed information about the candidates' lab experiences and how they handled setbacks.

Vogelstein and Kenneth Kinzler — who co-direct the Ludwig Center at Johns Hopkins in Baltimore, Maryland — run their group of about 15 trainees as subgroups of 2 or 3 people. The subgroups function as a risk-mitigation plan: all the members are co-authors on any papers. So even if one member has a particularly difficult project that takes three years to publish, she or he will have other papers come out in the interim.

And there are team rewards. Whenever work in the lab generates intellectual property, everyone in the lab benefits financially from any royalties. "Establishing a group that can cheer when someone else succeeds is not an accident," Vogelstein says. "It requires structure and planning of how people appreciate each other."

But even within a bonded community, trainees will get stuck at some point. Weekly check-ins or progress reports can prevent any hiding of problems. Lab heads can use software such as Trello or Slack, which allows everyone on a project to see progress — or lack thereof. Vogelstein has a pre-emptive approach. "Don't wait till they get stuck! The time to intervene is way before someone is despondent."

Switching a stuck lab member to a project



Anne Straube runs Cake Club in her department.

STEPHEN ROYLE

with better chances of success is a common strategy. But Stassun says that sometimes the best option is just to push through. “The single most important thing you can do is pull the bedcovers back and walk out the door,” he says. He uses ShareLaTeX, a shared online text-editing application, as a quick way to check for stalled project manuscripts. Stuck trainees should find one thing that they can write down, he says — maybe it’s one paragraph describing an experimental set-up or one paragraph of the introduction explaining a piece of background research.

Some senior scientists like to remind more-junior researchers that everyone gets mired at some point — it’s how they handle it that determines their success. When third-year PhD candidate Alice Bachmann, a member of Straube’s lab, got stuck for nearly 18 months on how best to prove that she had depleted a protein from her rat cells, she recalled the mantra of a friend: “If Plan A is not working, there are 26 letters in the alphabet.” Plan D ended up working, after she repeated it many times. Astrophysicist Rodolfo Montez Jr, a support scientist for the Chandra mission at the Smithsonian Astrophysical Observatory in Cambridge, Massachusetts, notes that he spends most of his research time running up against ‘bugs’ when computing an equation for an astronomical observation. “Solving the bug is now what your job is. It’s not a nuisance, it’s part of the path, and it’s cool,” he says.

That psychology of ‘flipping’ failure on its head is a recurring theme among lab leaders: embrace the failures, embrace the spinning wheels, embrace the bad weather at the field station. These things force researchers to get more creative and to approach problems in fresh ways. That, the leaders add, is when true discovery often happens.

Ultimately, the research enterprise works best when energy and enthusiasm remain high, even in the face of rejection, failure and defeat. “Keeping morale up in the lab is one of the most important aspects of trying to succeed,” says Vogelstein, whose group has identified more than a dozen major cancer genes, including the most common culprits in colon cancer.

“Succeeding in science is difficult,” he says. “We are always competing: for publications, for grants, for experiments. We are fighting battles all the time. The best thing to hear in the lab are the words, ‘It worked!’ You might think after 35 years it gets old, but it doesn’t.” ■

Kendall Powell is a freelance writer based in Lafayette, Colorado.

TURNING POINT

Nanotech bridge

Hossam Haick, a nanotechnology researcher at the Technion-Israel Institute of Technology in Haifa, has developed devices to detect cancer using exhaled breath rather than through biopsies. But, he explains, life in Israel can be difficult for Arab scientists. He is therefore trying to use his science to bridge cultural boundaries.

How many Arab professors are at the Technion?

Out of 600 faculty members, there are 9 Arab professors. The Arab community in Israel is 20% of the population, but in academia, it is roughly 1%. There is a pervasive belief in Israel that the Arab community is not educated. I try to dispel that notion.

How did you get the idea to use breath to diagnose disease?

I read a lot of the history. From the ancient Greeks 2,400 years ago to Alexander Graham Bell in the early 1900s, there were long-standing hypotheses of the smell of chronic disease in breath. I also heard hypotheses that dogs could smell cancer. I decided to see if I could prove scientifically whether there is something about exhaled breath that can reveal signs of disease.

At what stage is the research?

We have shown that exhaled breath contains unique fingerprints of specific diseases. We have lab results, as well as animal experiments. We have run clinical studies with 5,000 patients across 19 departments and 9 institutions, where we collected breath with a small device called NaNose, which is able to detect more than 1,000 different compounds in the breath from all of these people. We started with lung cancer, and have extended studies to gastric, colorectal and breast cancers, as well as to degenerative disease such as Parkinson’s, Alzheimer’s and multiple sclerosis. In the case of lung cancer, we are able to discriminate between benign and malignant tumours with 88% accuracy. By using breath to discriminate between benign and malignant tumours, we could save people from having to undergo unnecessary biopsies and surgeries.

Will this be in use soon?

We have built technology in a portable device that can detect disease in an easy and inexpensive way — only a few thousand dollars. Three companies have obtained the licence from Technion.

How else are you building on this technology?

We are also working on Sniffphone. The idea



is to bring breath analysers into smartphones. If a risk is found, the smartphone could send the results to a physician.

How do you try to improve relations between Arabs and Jews in Israel?

In a research institute, you are judged on excellence and achievements. That’s not the reality outside an academic institute, however. And in the Arab sector, there is a belief that whatever you do, you will not excel in Israel, unfortunately. As scientists, our role in the community is not only to produce papers. We have to disseminate the results and provide a message for the community. I volunteer to go to many community schools, both Jewish and Arab, to talk about science. These efforts consume a lot of my personal time. Every year, I give 200 lectures in schools. This is huge, considering I lead three research consortia. But it’s important.

How else do you pursue outreach?

A professor at the Technion had the idea that we should disseminate the fundamental principles behind nanotechnology and nanosensor research. We decided to present the course in a digital way, to talk to people beyond the boundaries of Israel. I said I would only do it if courses were offered in English and Hebrew, as well as in Arabic, so as not to discriminate. We have many people from Arab countries that have taken the course. One of the nicest data points is that 900 of those people who took my course were from Iran, which has huge political sensitivities with Israel. My students come from 127 countries worldwide, an indication that education, science and technology can serve for peace. ■

INTERVIEW BY VIRGINIA GEWIN

This interview has been edited for length and clarity.

WHEN THE COLD COMES

Be prepared.

BY DEBORAH WALKER

Before the funeral, I send seven sealed letters to the Unwalled Cities asking them to fulfil their obligations and send their best students to the Disease University and the Quarantine Security Academy. This is not the first request I've sent. I doubt the cities will honour their responsibilities, but the sealed letters will serve their purpose. Seven seals. If I were a Doctrinist, I might find some significance in that.

Although in Isolation Theta we are, in the main part, atheists, we appreciate ceremony as much as any Earth Doctrinist. The great and the good are gathered to attend the funeral of Dr Olinda Troy, the fourteenth Commander Pathologist of our colony.

Her coffin is empty. She's donated her body to the Disease U. I'm grateful. We're short of corpses, and my labs need bodies to test the latest antivirals. The Interferon-Zeconaril hybrid may, given time, prove effective.

When it's my turn to speak, I talk about an old Earth fable. The feckless grasshopper spent her summer playing the mandolin, while the ant toiled to build a store of food. Winter comes, as it always does, and in the cold the hungry grasshopper begs the ant for food and is refused. The food is for the ant and for her family. Hard work, sacrifice and planning are the ant's virtues.

"Dr Troy was an ant," I tell the congregation. "When the cold virus, Rhinovirus HRV-A488 mutated into the Bleeding Eyes serotype, she was ready. She activated the city walls. She marshalled the quarantine police. She enforced the daily testing regime, and expelled the infected: man and woman and child. She ensured our survival. Only through the harshest measures can we endure the Cold. And afterwards, she continued the fight, developing weapons against the myriad serotypes that wait for us. Through her will, she saved this colony.

As the fifteenth Commander Pathologist, I will do the same, if called to do so."

My eulogy is received in silence. There is a new ethos in the city. The next speaker tells it. "Compassion is the hallmark of this colony. We pity the scientists who genemodded the common-old viruses. We pity the doctors who undertook the clinical trial. We pity the lab workers who manufactured

and genemodded to have fatal potentiality. Repeated mutations, owing to low-fidelity replication and frequent recombination, mean that at any time antigenic shift may occur, and a new, fatal pandemic may arise. Witness the previous epidemics: Bleeding Eyes, The Judders, Heart Halt, Dry Dust Brain, The Weeping Trembles.

We cannot vaccinate against 500 constantly mutating strains, but I believe the antiviral hybrids will eventually generate a cure.

The grasshoppers, however, want to divert our resources to other things. Wouldn't it be fine to build another ship, they argue? To go off planet.

Of the walled cities we built before the attack, only Theta stands. Our colony ship was stationed at Epsilon. In the chaos of our first pandemic, the citizens of Epsilon fled the planet. But they took the Cold aboard, virions harboured in the air they breathed.

We never heard from the colony ship. Why would a new ship be any different?

I walk slowly to the labs. Sometimes, and this is a fanciful conceit, I imagine the

Cold as an entity. A spectre that's staying its hand. It knows that if it attacks in 20 years, it will find the cities overflowing with grasshoppers, full of pity, unwilling to do what is necessary. The Cold tires of the contest and in a few decades it will have its endgame.

A fancy only. A virus has no will. There was only the will of the Doctrinists who created it and set out to destroy us non-believers. And there is only my will that will keep the game in play.

I have sent seven sealed letters. When the seals are opened, a new pandemic will be born. The Cold will come, killing grasshoppers and ants alike. But I will be here. I will ensure that we endure. I will set a new generation of ants working towards the cure.

I am the sender of the seals. I am the har-binger of the plague. I am the doctor who sends Cold into our world.

Do not pity me. ■

Find Deborah in the British Museum trawling the past for future inspiration.

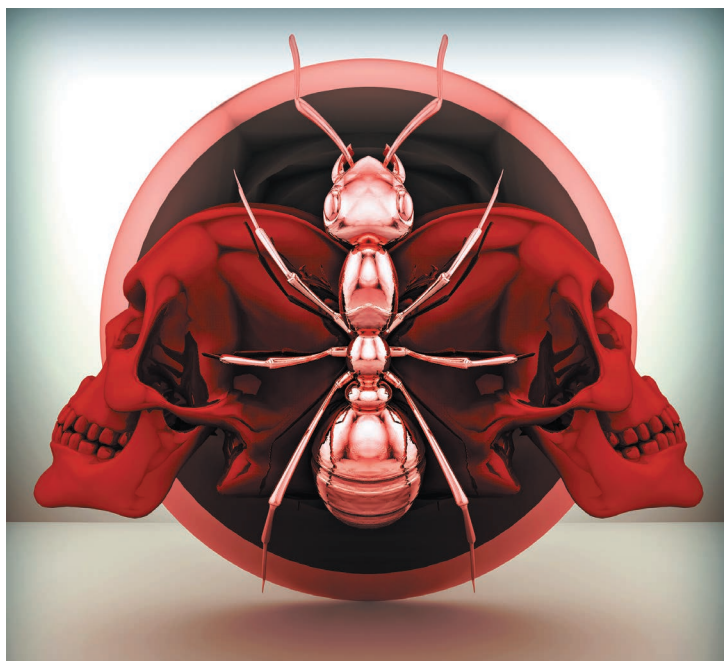


ILLUSTRATION BY JACEY

the viruses. We pity the physicists who plotted the course of the bioweapon. We pity the Doctrinist Senate that approved the decision. We pity every Doctrinist man and woman who condoned that act to send the bioweapon hurtling after the people fleeing Earth. We pity them all."

It's a grasshopper sentiment, as Dr Troy often remarked. She's dead. She'll not have to witness another epidemic. There is my pity.

The service ends, the grasshoppers move out quickly. We ants are slower, older.

I pass two grasshoppers. The woman clings to her husband. "I can't believe Earth sent the Cold after us. What hatred they had."

"I pity them," says the man sanctimoniously.

Pity is a fine thing. But do the grasshoppers think it will save them? They're lulled because the Cold has been quiet for half a century. Yet there is no vaccination against the Cold, and little to no cross-protection against the serotypes. Earth sent us at least 500 serotypes, highly contagious